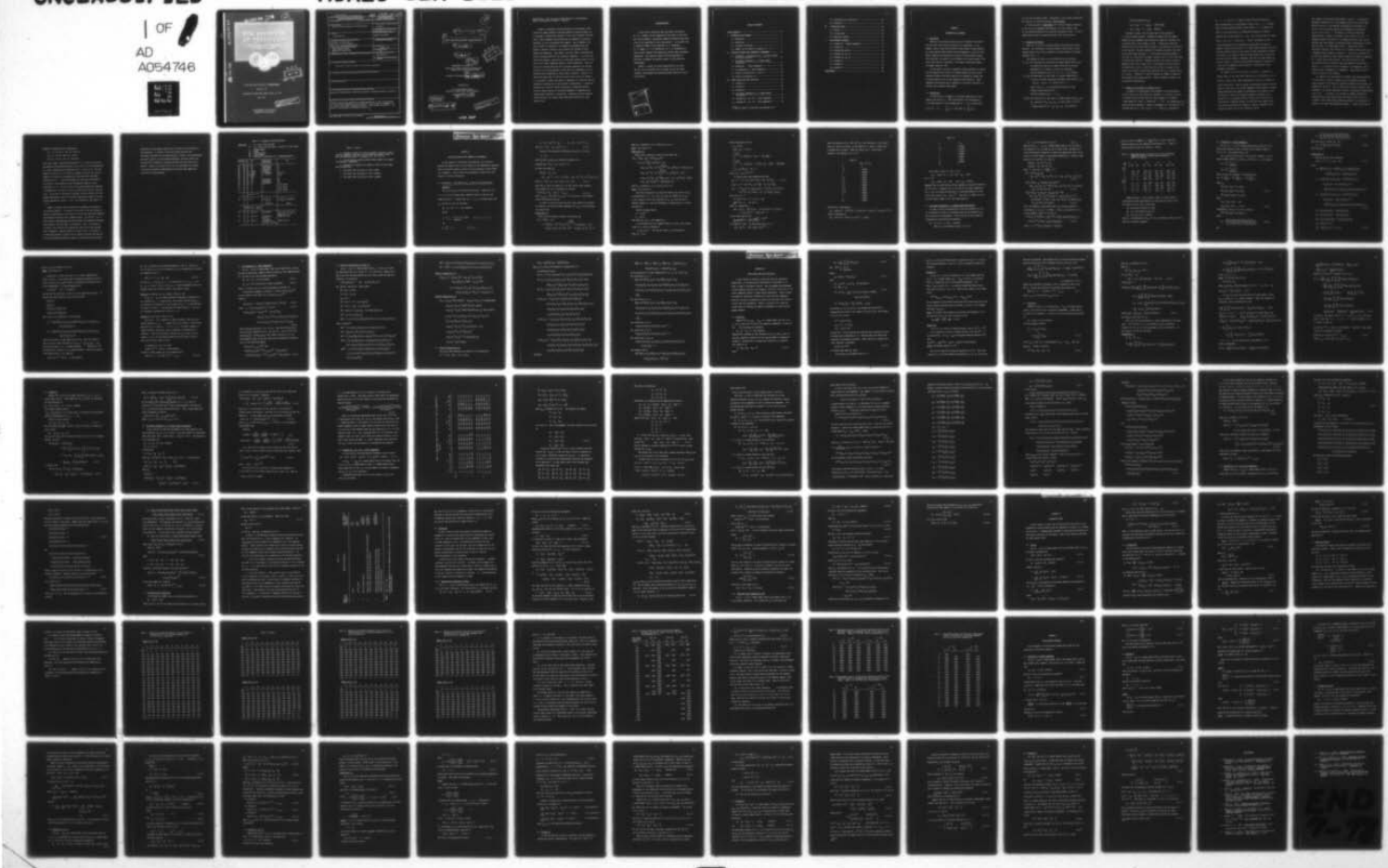
AD-A054 746 NORTH CAROLINA UNIV AT CHAPEL HILL INST OF STATISTICS F/G 12/1
TESTS FOR FINITE MIXTURES OF DISTRIBUTIONS.(U)
APR 77 C HSU DAAG29-74-C-0030
UNCLASSIFIED MIMEO SER-1122 ARO-11959.20-M NL



SIFIED

1 OF

AD

A054746

THE INSTITUTE OF STATISTICS

THE CONSOLIDATED UNIVERSITY
OF NORTH CAROLINA

TESTS FOR FINITE MIXTURES OF DISTRIBUTIONS

Chin-Fei Hsu

Institute of Statistics Mimeo Series No. 1122

May, 1977

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ARO 11959:20-M

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Tests for Finite Mixtures of Distributions		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL
		6. PERFORMING ORG. REPORT NUMBER Mimeo Series No. 1122
7. AUTHOR(s) Chin-Fei Hsu		8. CONTRACT OR GRANT NUMBER(s) DAAG29-74-C-0030
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics University of North Carolina Chapel Hill, North Carolina 27514		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Army Research Office Research Triangle Park N.C.		12. REPORT DATE April 1977 13. NUMBER OF PAGES 89
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for Public Release: Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Mixtures of distributions, rank-type statistics, Kolmogorov-Smirnov statistic.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Derive and discuss properties of three approaches to test the hypothesis of finite proper mixtures of distributions. Two of them are compared for the case of two normal components against a single normal alternative. The third is investigated by simulation studies.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(2)

(9) Technical Rept.

(6)

TESTS FOR FINITE MIXTURES OF DISTRIBUTIONS

(10) Chin-Fei Hsu

(11) April 1977

(12) 96 p.

Department of Statistics
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina 27514

(14)

Institute of Statistics Mimeo Series No. 1122

Sponsored by

U.S. Army Research Office
Research Triangle Park, North Carolina

Contract DAAG29-74-C-0030

(15)

(18) ARO

(19) 11959.20-M

410 064

AB

CHIN-FEI HSU. Tests for Finite Proper Mixtures of Distributions (Under the direction of Norman L. Johnson)

A number of hypothesis testing problems are investigated, which involve the common element of deciding whether an observed sample can be regarded as coming from a mixture of two or more component distributions. These component distributions may be completely or partially specified, or only be estimated from samples. One of Johnson's statistics (1973) is studied for its asymptotic performance and this method is applied to derive a test statistic for mixture of three symmetrical components. Next Thomas' statistic (1969) is modified and used in testing mixtures of two continuous components. It is then compared with Johnson's statistics by calculating asymptotic power for the case of two normal components against a single alternative. Then statistics for testing three and four continuous components are derived. Statistics for testing mixtures of more than four components are also discussed and an algorithm to derive them is obtained. Further it is shown that these kinds of statistics can be used to test (i) whether it is possible to reduce the number of components in a mixture, and (ii) hypotheses involving two or more mixtures simultaneously. A method of deriving test statistics based on minimizing a Kolmogorov-Smirnov statistic among mixtures of two known components is suggested and a computational algorithm is constructed. Properties of tests based on these statistics are studied using simulation procedures for some special cases.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my advisor, Dr. N. L. Johnson, for his suggestion on the topic of this work. His overall guidence, inspiration, numerous discussions have been vital to the completion of this long project. I also would like to thank the members of my committee, Dr. W. Hoeffding, Dr. G. Simons, Dr. I. M. Chakravarti, Dr. C. M. Suchindran for their careful reading of the manuscript and for their invaluable suggestions. The financial support from the Department of Statistics throughout my graduate studies is also gratefully acknowledged.

Finally, I express my sincere appreciation to my wife, Chi-Yin, for her excellent job of typing, and for her understanding, encouragement and enduring patience during the course of this work.

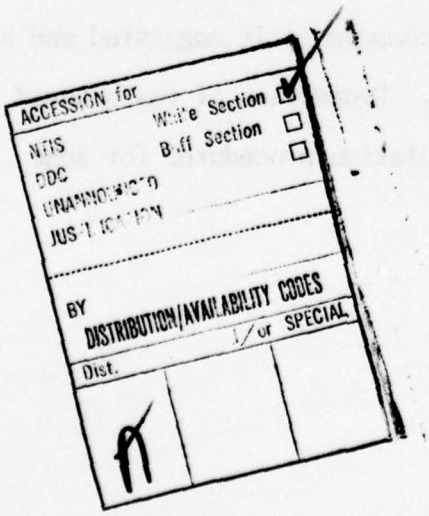


TABLE OF CONTENTS

[1] A table on pages 8-9 describes the problems A1-F3

CHAPTER I

INTRODUCTION AND SUMMARY

1.1 Motivation

Mixtures of distribution functions arise frequently in practice and very often they present difficulites to researchers. As an example, similar items from different sources might be mixed together at a distribution center before they are shipped out to lots for sale. Doubts regarding the uniformity of quality of these products having been expressed, a procedure to test whether lots contain products from two or more sources is desirable. For another interesting example, see Thomas (1969, pp. 475).

In this paper a variety of hypothesis testing problems (or models) are investigated which possess the common element that they involve questions whether an observed sample can be regarded as coming from a finite mixture of two or more component (distributions). These component distributions may be completely or partially specified, or possibly only estimated from samples.

1.2 Definitions

A finite mixture is proper if its mixing coefficients are non-nagative and sum up to 1. More specifically, the distribution F_0 is a proper mixture of the distributions F_1, ---, F_k if there exist

$$\omega_i \geq 0,\ i=1,---,k, \qquad \sum_{i=1}^{k} \omega_i = 1 \text{ such that } F_0 = \sum_{i=1}^{k} \omega_i F_i ,$$

ω_i's are not necessarily known. Furthermore, if the mixing coefficients are positive, we called the mixture strictly proper.

A finite mixture is identifiable if it can be uniqurely expressed as far as (1) the component distributions, (2) their numbers, and (3) the mixing coefficients are concerned (Behboodian, 1976). A finite identifiable mixture is necessarily proper, but not vice versa.

1.3 Grouping of Problems

Table 1.1 is a list of testing problems concerning finite proper mixtures. For convenience, relevant problems, which either (1) have similar nature, or(2) can be treated by a similar method, are grouped together into A - F.

The contents of table 1.1 are illustrated by the following:

i). A2: The conditions are that there are random samples from F_0 and F_1 respectively and all the random variables are mutually independent; F_2 is known; F_1 and F_2 are continuous.
The null hypothesis H_0 is $F_0 = \omega F_1 + (1 - \omega)F_2$ properly.

ii). B3: The conditions are that there is a random sample from F_0; F_1 and F_2 are known, absolutely continuous with density f_1, f_2;
$f_0(x,\omega) = \omega f_1(x) + (1 - \omega)f_2(x)$; for $\omega' > \omega$
$f_0(x,\omega')/f_0(x,\omega)$ is a nondecreasing function of some suitably chosen function $t(x)$.
The null hypothesis H_0 is $\omega \leq \omega_0$.

iii). F2: The conditions are that there are random samples from F_{a0} and F_{b0}, and each of F_{a1}, F_{a2}, F_{b2} is either known or there is a random sample for it; F_{a1}, F_{a2}, F_{b2} are continuous.

The null hypothesis H_0 is

$$F_{a0} = \omega_a F_{a1} + (1 - \omega_a)F_{a2} \quad \text{properly and}$$

$$F_{b0} = \omega_b F_{a1} + (1 - \omega_b)F_{b2} \quad \text{properly.}$$

Problems in group A test the same kind of null hypotheses, i.e. that of proper mixtures. Problems in group B have the common element that the component distributions are symmetric and differ only in location shift, scale change, or both. Though these conditions seem more restrictive than those in group A, the condition of continuity is not required. The null hypotheses of problems in group C deal with "restricted" proper mixtures, i. e. the mixing proportion parameter ω lies only in a restricted (or proper) interval [a,b] of [0,1] other than the entire unit interval. Problems in group D are similar to those of goodness-of-fit tests, except that (i) the sample(s) is(are) from the distribution(s) (F_0 or F_1) other than the one (f_2) we intend to test, and(ii) we are given an additional condition that F_0 is a proper mixture of F_1 and F_2. Problem E1 is that of reducing the number of components from a given finite proper mixture. Finally, problems in group F test two mixtures simultaneously.

1.4 Summary of the Results in Chapter II to V

In chapter II we deal mainly with problems of location mixtures. First we derive the third and fourth cumulants of a statistic proposed by Johnson (1973) and find that this third cumulant is a constant (with respect to ω and n) multiple of $n^{-.5}(1 - 2\omega)$ (Proposition 2.1), while the fourth cumulant is entirely independent of ω (Proposition 2.2), provided that $\Pr\{ X_1 = (m_1 + m_2)/2 \} = 0$, where m_i is the mean of

F_i, $i = 1, 2$, and X_1 is a random variable having distribution F_0. Next assuming that F_i is distributed as $N(m_i,\sigma^2)$ for $i = 1, 2$, (where $N(m_i,\sigma^2)$ means normal with mean m_i and variance σ^2,) we compute the third (γ_1) and the fourth (γ_2) standardized cumulants of Johnson's statistic for various values of $\Delta = (m_2 - m_1)/\sigma$ and find that for sample size $n \geq 100$, $|\gamma_1| < .029$ and $|\gamma_2| < .01131$. Furthermore we use a 4-term Gram-Charlier series expansion to approximate the power of Johnson's statistic with respect to a single normal alternative and find that the values of power are very close to those calculated by Johnson. In section 2.3 we propose a statistic which extends that of Johnson to the case of three symmetric components, and derive a large sample test. Next in section 2.4 we derive an approximate formula for the power of a test for tow components mixture against a "proportional" three components alternative.

In chapter III we first describe a statistic T_2, proposed by Thomas (1969), in the form used by Hariton (1972) and modify it (to T_3 and T_4) to obtain large sample (nonparametric) tests for problems A2 and A3 respectively. Then we compute the approximate power of the test using T_4 in the case of two normal components mixture against a single normal alternative and compare its performance with Johnson's tests. In sections 3.5 and 3.6 we extend the method of deriving a test statistic for two component mixture to obtain test statistics for three and four component mixtures and find that large sample tests can easily be formulated for problem A3--when all component distributions are known - but not for problems A1 and A2.

This suggests that we might modify Thomas' statistic. By regarding a homogeneous population as a "one component mixture" and using a Mann-Whitney statistic to test whether the population distribution is a specific one, the testing problem can be embedded into problem A1 (or A2, or A3) with k = 1. By doing so, we find an interesting algorithm which enables us to obtain a test statistic for a (k + 1) components mixture from that for a k components mixture, when k=1,2,3, provided that each component distribution under consideration is either known or there is a random sample from it. Application of this algorithm to derive a test statistic for 5-component mixture from that of 4-component mixture appears in section 3.7 In the rest of Chapter III we deal with problem E1, which involves reducing the number of components for a given finite proper mixture. Two special cases are fully investigated under the condition that each component distribution is known, i. e. (1) reduction from 4 components to 3 components, and (2) reduction from 3 components to 2 components. Large sample tests for these two cases are also obtained.

In chapter IV we investigate by the Monte Carlo method properties of a statistic D, which we propose for use in problems involving proper mixtures of two known (or specified) component distributions, e.g. problems A3, B1 and D2. First we define this statistic formally, then rearrange it in another form which is easier to manipulate computationally (see 4.1.4). In section 4.2 we describe a computer algorithm to calculate from a given random sample the values of both D and $\hat{\omega}$, an estimator of the actual mixing proportion parameter in the mixture. In section 4.3 we study three numerical examples —— namely, when the two

component distributions are respectively:

(i) $F_1 \sim N(-1.5,1)$ and $F_2 \sim N(1.5,1)$

(ii) $F_1 \sim N(-2,1)$ and $F_2 \sim N(2,1)$

(iii) $F_1 \sim E(1.5)$ and $F_2 \sim N(1,1/16)$,

where "E(a)" means "exponential with mean a^{-1}". In order to calculate the empirical distribution function (e.d.f.) of D, it is first expressed as a linear combination of conditional e.d.f.'s (see 4.1.6). Then 200 samples of sizes 10, 20, 40, each for example (i) and 10, 20, each for examples (ii) and (iii) are generated to calculate each individual conditional e.d.f.. From these examples we find that the average of the values of the estimators $\hat{\omega}$ over 200 generate samples in each conditional e. d. f. are very close to the actual proportions parameter ω, provided that ω are not too close to either end of the unit interval [0,1]. To see how the statistic D performs, we also calculate for each example approximate values of the actual significance levels of D using nominal significance levels α = .05, .025 respectively (see Tables 4.5 — 4.7).

Chapter V contains short discussions of problems not included in the previous chapters. In section 5.1 we derive a test for problem B2, using the sample mean as test statistic, for the case when both component distributions are normal with a common variance. In section 5.2, we first extend a lemma dealing with distributions with monotone likelihood ratio property, and then apply it to mixtures. Next, in section 5.3, we derive a test statistic for problem B4, which can be used together with a Kolmogorov - Smirnov statistic to form a test. In section 5.4, we show that problems in group C can be treated in exactly the same way as the corresponding problems in group A by proving that the necessary

conditions in both groups, which we use to derive the test statistics, are equivalent. In section 5.5 we first obtain necessary and sufficient conditions for the function $F_0(x) - \omega F_1(x)$ to be nondecreasing when both F_0 and F_1 are distribution functions, and then propose test statistics for problems D1, D2 and D3 (without deriving their distribution). Finally in sections 5.6 to 5.8, we investigate the problems of testing two mixtures simultaneously and derive large sample test statistics for such problems.

Table 1.1 Grouping of Testing Problems

(Notations: * : $F_0 = \omega F_1 + (1-\omega)F_2$ properly
** : $F_0 = \omega F_1 + (1-\omega)F_2$ for some ω, $0<a\le\omega\le b<1$, a and b known
S : random sample
K : known
LS : location shift
SC : scale change
AC : absolutely continuous w.r.t. Lebesgue measure.)

Symbol	F_0	F_1	F_2	Additional conditions	H_0
A1[1]	S^2	S^2	S^2	continuity[3]	*
A2	S	S	K	continuity	*
A3	S	K	K	continuity	*
B1	S	K	K,LS[4]	symmetry[3]	*
B2	S	K	K,LS	*, symmetry	$\omega=0$, or 1
B3	S	K,AC	K,AC	*, (5.2.5)	$\omega\le\omega_0$
B4	S	K	LS[5]	-	*
B5	S	K	SC[6]	-	*
C1	S	S	S	continuity	**
C2	S	S	K	continuity	**
C3	S	K	K	continuity	**
D1	S	S	-	*	$F_2=F$, known
D2	S	K	-	*	$F_2=F$, known
D3	K	S	-	*	$F_2=F$, known
		F_i, i=1,---,k			
E1	S	S or K		$F_0=\sum_{i=1}^{k}\omega_i F_i$ properly	$F_0=\sum_{i=1}^{m}\omega_i F_i$ properly, $2\le m<k$
	F_{a0}	F_{b0}	F_{ij}, i=a,b; j=1,2		
F1	S	S	S or K	continuity	$F_{i0}=\omega_i F_{i1}+(1-\omega_i)F_{i2}$ properly, i=a,b
F2	S	S	S or K	continuity, $F_{a1}=F_{b1}$	same as in F1
F3	S	S	S or K	continuity, $F_{a1}=F_{b1}$, $F_{a2}=F_{b2}$	same as in F1

Table 1.1 (Cont.)

1. Problems in groups A--D can be extended to mixtures of more than two components, and they will be denoted by adding a 'prime' to the corresponding symbols, for examples, A1', B2', C3' etc.

2. In each problem, random variables among samples are assumed to be mutually independent.

3. Continuity or symmetry are assumed to hold for every component distributions.

4. This means that $F_2(x)=F_1(x-t)$ with t known.

5. This means that $F_2(x)=F_1(x-t)$ with t unknown.

6. This means that $F_2(x)=F_1(x/t)$ with t unknown.

CHAPTER II

LOCATION MIXTURES WITH SYMMETRICAL COMPONENTS

In this chapter we investigate the properties of a statistic proposed by Johnson (1973) to test mixture of two symmetrical components. Then we extend the method to derive a test for mixture of three symmetric components. Also we study the performance of these tests with respect to certain alternatives.

2.1 Problem B1 — the Statistic $\hat{\omega}_x - \hat{\omega}_y$ and Its Third and Fourth Cumulants

Let F_0, F_1 and F_2 be distribution functions. Assume that, for $i = 1, 2$, F_i are of known forms, symmetric about its mean m_i with common variance σ^2. Suppose that $X_1, ---, X_n$ is a random sample from F_0. We like to test the following:

H_0: $F_0 = \omega F_1 + (1 - \omega)F_2$ properly.

Define

$$Y_i = \begin{cases} 1 & \text{if } X_i < (m_1 + m_2)/2 \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } i = 1, 2, ---, n.$$

Let

$$\bar{X} = \frac{1}{n} \Sigma X_i , \qquad \bar{Y} = \frac{1}{n} \Sigma Y_i ,$$

$$\hat{\omega}_x = (m_1 - m_2)^{-1}(\bar{X} - m_2), \qquad \hat{\omega}_y = (P_1 - P_2)^{-1}(\bar{Y} - P_2),$$

where $P_i = \Pr\{X_1 < (m_1 + m_2)/2 \mid F_i\}$. (2.1.1)

Johnson (1973) proposes the following statistic to test H_0:

$$\hat{\omega}_x - \hat{\omega}_y .$$

Under H_0 both $\hat{\omega}_x$ and $\hat{\omega}_y$ are unbiased estimators of ω.

Assuming that $\Pr\{X_1 = (m_1 + m_2)/2\} = 0$,

he shows that under H_0 ,

$$n\mathrm{Var}(\hat{\omega}_x - \hat{\omega}_y) = n\kappa_2$$

$$= (m_1 - m_2)^{-2}\sigma^2 + (P_1 - P_2)^{-2}P_1P_2 - 2(m_1 - m_2)^{-1}(P_1 - P_2)^{-1}P_1(E_1 - m_1) ,$$

where $E_i = E(X_1 \mid X_1 < (m_1 + m_2)/2, F_i)$, $i=1,2$. (2.1.2)

Note that $n\kappa_2$ does not depend on ω. By the central limit theorem a large sample test can be formulated as follows:

$$\text{reject } H_0 \text{ if } |\hat{\omega}_x - \hat{\omega}_y|\kappa_2^{-\frac{1}{2}} > z_{1-\alpha/2} , \qquad (2.1.3)$$

where $z_{1-\alpha/2}$ is such that $\Phi(z_{1-\alpha/2}) = 1-\alpha/2$ and $\Phi(x)$ is the standard normal distribution function.

In order to see more clearly how this large sample test performs we calculate the third and fourth cumulants of $\hat{\omega}_x - \hat{\omega}_y$ in the following two propositions.

Proposition 2.1

With $\hat{\omega}_x$ and $\hat{\omega}_y$ defined as above, we have under H_0,

$$n^2\kappa_3 = n^2E(\hat{\omega}_x - \hat{\omega}_y)^3$$

$$= (1-2\omega)\Big\{3(m_1 - m_2)^{-2}(P_1-P_2)^{-1}\Big[\int_{-\infty}^{\frac{m_1+m_2}{2}} (x-m_1)^2 dF_1(x) - P_1\sigma^2\Big]$$

$$+ 3P_1(E_1 - m_1)(m_1 - m_2)^{-1}(P_1 - P_2)^{-1} - P_1P_2(P_1 - P_2)^{-2}\Big\} \qquad (2.1.4)$$

where P_i is defined by (2.1.1) and E_1 by (2.1.2).

Proof: See section 2.5.

Proposition 2.2

With $\hat{\omega}_x$ and $\hat{\omega}_y$ defined as above, we have under H_0,

$$n^3\kappa_4 = n^3E(\hat{\omega}_x - \hat{\omega}_y)^4 - 3n^3[E(\hat{\omega}_x - \hat{\omega}_y)^2]^2$$

$$= (m_1 - m_2)^{-4}\left[\int_{-\infty}^{\infty}(x - m_1)^4 dF_1(x) - 3\sigma^4\right]$$

$$-4(m_1 - m_2)^{-3}(P_1 - P_2)^{-1}\left[\int_{-\infty}^{\frac{m_1+m_2}{2}}(x - m_1)^3 dF_1(x) - 3\sigma^2P_1(E_1 - m_1)\right]$$

$$-6(m_1 - m_2)^{-2}(P_1 - P_2)^{-1}\left[\int_{-\infty}^{\frac{m_1+m_2}{2}}(x - m_1)^2 dF_1(x) - P_1\sigma^2\right]$$

$$-12(m_1 - m_2)^{-2}(P_1 - P_2)^{-2}P_1^2(E_1 - m_1)^2 + (P_1 - P_2)^{-4}P_1P_2(1 - 6P_1P_2)$$

$$-4(m_1 - m_2)^{-1}(P_1 - P_2)^{-2}(1 - 6P_1P_2)P_1(E_1 - m_1),$$

where P_i is defined by (2.1.1) and E_1 by (2.1.2).

Proof: See section 2.5.

From these propositions we find that under H_0 the values of $n^2\kappa_3$ is proportional to $(1 - 2\omega)$, while $n^3\kappa_4$ does not depend on ω at all. In all, among the first four cumulants of $\hat{\omega}_x - \hat{\omega}_y$, only the third cumulant depends on ω and this dependence is proportional to a linear function of ω.

Define the shape factors

$$\gamma_1 = \kappa_3/\kappa_2^{3/2},$$

$$\gamma_2 = \kappa_4/\kappa_2^2.$$

Then under H_0, only γ_1 will depend on ω.

In the following we will compute values of γ_1 and γ_2 for various values of Δ, which is defined as

$\Delta = (m_2 - m_1)\sigma^{-1}$, for the case that F_i is distributed as $N(m_i, \sigma^2)$, $i=1,2$.

Then by definition we have

$$P_1 = \Phi(\Delta/2) ,$$

$$P_1(E_1 - m_1) = -\sigma z(\Delta/2) ,$$

$$\int_{-\infty}^{\frac{m_1+m_2}{2}} (x - m_1)^2 dF_1(x) - P_1\sigma^2 = -\frac{\Delta}{2} z(\tfrac{\Delta}{2})\sigma^2 ,$$

$$\int_{-\infty}^{\frac{m_1+m_2}{2}} (x - m_1)^3 dF_1(x) - 3\sigma^2 P_1(E_1 - m_1) = z(\tfrac{\Delta}{2})\sigma^3 - (\tfrac{\Delta}{2})^2 z(\tfrac{\Delta}{2})\sigma^3 ,$$

$$\int_{-\infty}^{\infty} (x - m_1)^4 dF_1(x) - 3\sigma^4 = 0 ,$$

where $z(x) = (2\pi)^{-1/2} e^{-x^2/2}$.

It follows after some simplification that

$$n\kappa_2 = \Delta^{-2} - 2\Delta^{-1}(P_1 - P_2)^{-1} z(\Delta/2) + (P_1 - P_2)^{-2} P_1 P_2 , \qquad (2.1.5)$$

$$n^2\kappa_3(1 - 2\omega)^{-1} = \tfrac{3}{2}\Delta^{-1}(P_1 - P_2)^{-1} z(\tfrac{\Delta}{2}) - (P_1 - P_2)^{-2} P_1 P_2$$

$$n^3\kappa_4 = -\Delta^{-1}(P_1 - P_2)^{-3}(2 - 16P_1P_2) z(\Delta/2) - 12\Delta^{-2}(P_1 - P_2)^{-2} z^2(\Delta/2)$$
$$+ 4\Delta^{-3}(P_1 - P_2)^{-1} z(\Delta/2) + (P_1 - P_2)^{-4} P_1 P_2 (1 - 6P_1P_2) .$$

Note that as $x \to \infty$,

$$[1 - \Phi(x)] z^{-1}(x) = x^{-1} + o(x^{-1}), \text{ and}$$

$$\lim_{x \to \infty} x^k z(x) = 0 , \text{ for any } k .$$

Hence as Δ approaches ∞ ,

$$\sqrt{n}(1 - 2\omega)^{-1}\gamma_1 \sim [\tfrac{3}{2}\Delta^{-1} z(\Delta/2) - 2\Delta^{-1} z(\Delta/2)][\Delta^{-2} - 2\Delta^{-1} z(\Delta/2)$$
$$+ 2\Delta^{-1} z(\Delta/2)]^{-3/2} \sim -\tfrac{1}{2}\Delta^2 z(\Delta/2) \to 0 .$$

On the other hand we have

$$\lim_{\Delta \to 0} \frac{\Delta}{2\Phi(\Delta/2) - 1} = \lim_{\Delta \to 0} \frac{(2\pi)^{1/2}}{e^{-\Delta^2/8}} = (2\pi)^{1/2}$$

by L'Hôpital's rule. It then follows that as $\Delta \to 0$

$$\sqrt{n}(1 - 2\omega)^{-1}\gamma_1 \sim \Delta(\tfrac{3}{2} - \tfrac{\pi}{2})[1 - 2 + \tfrac{\pi}{2}]^{-3/2} \to 0.$$

Since the function (of Δ), $\sqrt{n}(1-2\omega)^{-1}\gamma_1$ is not identical to zero, there exists at least one extreme in the range $(0,\infty)$. Table 2.1 shows that a minimum value of about -.28961 for $\sqrt{n}\gamma_1(1-2\omega)^{-1}$ occurs when Δ belongs to the interval (2.71, 2.73).

Table 2.1

Δ	$\sqrt{n}(1-2\omega)^{-1}\gamma_1$
.0	.0
.01	-.001642
.1	-.016410
1	-.15681
1.5	-.22022
2.0	-.26510
2.5	-.28735
2.7	-.28959
2.71	-.28961
2.72	-.28961
2.73	-.28961
3	-.28604
5	-.12905
10	-.00006

Similarly as Δ approaches ∞,

$$n\gamma_2 \sim [4\Delta z(\Delta/2) - 2\Delta^2 z^2(\Delta/2) + 2\Delta^3 z(\Delta/2)][1- 2z(\Delta/2)\Delta^{-1} + 2z(\Delta/2)\Delta^{-1}]^{-2} \to 0$$

and as Δ approaches 0,

$$n\gamma_2 \sim [4-12+ 4\pi- \pi^2/2][1- 2+ \pi/2]^{-1} = -1.13082$$

Table 2.2

Δ	$n\gamma_2$
0	-1.13082
.01	-1.13081
.1	-1.12938
.5	-1.09608
1	-1.00679
5	-.16626
10	-.000047
15	0

From tables 2.1 and 2.2, for $n \geq 100$

$$|\gamma_1| < .029|1- 2\omega| \leq .029 \text{ , and } |\gamma_2| < .01131 .$$

For $n \geq 25$,

$$|\gamma_1| < .058|1- 2\omega| \leq .058 \quad \text{and} \quad |\gamma_2| < .045233 .$$

Remember that the third and the fourth cumulants of normal distributions are both zero. It appears that in the case of two normal components, for $n \geq 100$, the normal distribution is a good approximation to that of the test statistic $(\hat{\omega}_x - \hat{\omega}_y)\kappa_2^{-1/2}$. Even for n as small as 25, we can still expect normal to be a good approximation.

2.2 Two Normal Components vs. a Single Normal Alternative

In order to assess the properties of the test procedure (2.1.3), it is desirable to study its performance when the distribution F_0 is not a mixture of two components. In this section we will study the following null and alternative hypotheses:

H_0: $F_0 = \omega F_1 + (1- \omega)F_2$ properly,

where F_i is distributed as $N(m_i, \sigma^2)$, i=1,2.

H_1: F_0 is distributed as $N(\mu,\tau^2)$. (2.2.1)

Suppose that X_1, ---, X_n is a random sample from F_0. We will derive the cumulants of $\hat{\omega}_x - \hat{\omega}_y$ under H_1, up to fourth order, and then use a 4-term Gram-Charlier expansion to compute the approximate power of the test (2.1.3) with respect to the above alternative H_1. Using the same notation as in section 2.1, under H_1

$$E(\hat{\omega}_x|H_1) = (m_1 - m_2)^{-1}(\mu - m_2)$$

$$E(\hat{\omega}_y|H_1) = (P_1 - P_2)^{-1}(P - P_1) ,$$

where $P = \Phi(K)$ and $K = (\frac{m_1+m_2}{2} - \mu)\tau^{-1}$,

$$nV_2 = nVar(\hat{\omega}_x - \hat{\omega}_y | H_1)$$

$$= (m_1 - m_2)^{-2}\tau^2 + (P_1 - P_2)^{-2}P(1-P) + 2(m_1 - m_2)^{-1}(P_1 - P_2)^{-1}\tau z(K) \quad (2.2.2)$$

$$n^2V_3 = n^2E[\hat{\omega}_x - \hat{\omega}_y - E(\hat{\omega}_x|H_1) - E(\hat{\omega}_y|H_1)|H_1]^3$$

$$= 3(m_1 - m_2)^{-2}(P_1 - P_2)^{-1}\tau^2 Kz(K) - 3(m_1 - m_2)^{-1}(P_1 - P_2)^{-2}(1-2P)\tau z(K)$$

$$- (P_1 - P_2)^{-3}P(1-P)(1-2P) \quad (2.2.3)$$

$$n^3V_4 = n^3\{E[\hat{\omega}_x - \hat{\omega}_y - E(\hat{\omega}_x|H_1) - E(\hat{\omega}_y|H_1)|H_1]^4 - 3V_2^2\}$$

$$= 4(m_1 - m_2)^{-3}(P_1 - P_2)^{-1}\tau^3 z(K)(K^2-1) - 6(m_1 - m_2)^{-2}(P_1 - P_2)^{-2}\tau^2 \cdot$$

$$[(1-2P)Kz(K) + 2z^2(K)] + 4(m_1 - m_2)^{-1}(P_1 - P_2)^{-3}(1-6P+6P^2)\tau z(K)$$

$$+ (P_1 - P_2)^{-4}P(1-P)(1-6P+6P^2) \quad (2.2.4)$$

For a 5% asymptotic significance level, the critical region for rejecting H_0 is $\{|\hat{\omega}_x - \hat{\omega}_y| > 1.96\sqrt{\kappa_2}\}$, where κ_2 is given by (2.1.5). So the asymptotic power of the test is

$$Pr\{|\hat{\omega}_x - \hat{\omega}_y| > 1.96\sqrt{\kappa_2}|H_1\} \simeq F_\xi(V_2^{-1/2}[-1.96\sqrt{\kappa_2} - E(\hat{\omega}_x|H_1) + E(\hat{\omega}_y|H_1)])$$

$$+ 1 - F_\xi(V_2^{-1/2}[1.96\sqrt{\kappa_2} - E(\hat{\omega}_x|H_1) + E(\hat{\omega}_y|H_1)]) \quad (2.2.5)$$

where $\xi = V_2^{-1/2}[\hat{\omega}_x - \hat{\omega}_y - E(\hat{\omega}_x|H_1) + E(\hat{\omega}_y|H_1)]$,

and $F_\xi(x) = \Phi(x) - z(x)[\frac{1}{6}V_3(x^2-1) + \frac{1}{24}V_4(x^3-x)]$ is the 4-term Gram-Charlier expansion of the cumulative distribution of ξ. Table 2.3 contains values of power for various τ, μ and m_1.

Table 2.3 Approximate power of the test (2.1.3) for two normal components against a single normal alternative (α=5%, σ =1, m_2=-m_1)

$\frac{\lvert\mu\rvert}{m_1}$	τ=	.5		1.0		2.0		5.0	
	n=	100	400	100	400	100	400	100	400
m_1=1									
.2		.799	.999	.0786	.130	.265	.501	.693	.753
.4		.999	1.0	.109	.262	.505	.915	.753	.895
.6		1.0	1.0	.116	.305	.758	.997	.827	.975
.8		1.0	1.0	.0825	.180	.918	1.0	.895	.997
m_1=2									
.2		1.0	1.0	.674	.989	.201	.347	.685	.853
.4		1.0	1.0	.962	1.0	.385	.789	.853	.993
.6		1.0	1.0	.981	1.0	.679	.989	.958	1.0
.8		1.0	1.0	.778	1.0	.920	1.0	.993	1.0

Comparing Table 2.3 with Johnson's Table Ib (1973), which is calculated by normal approximation, we find that the values in the two tables are very close to each other.

For the reverse situation in which

H_0: F_0 is distributed as $N(\mu,\sigma^2)$,

H_1: $F_0 = \omega F_1 + (1-\omega)F_2$ properly,

where F_i is distributed as $N(\mu_i,\sigma^2)$ i=1,2, see Bryant (1973, pp. 28-37), who uses a different approach.

2.3 Problem B1' — Three Components

Let $X_1, X_2, \cdots, X_n$ be a random sample from F_0 and $F_i(x)$ be of known forms, symmetric about its mean m_i with common variance σ^2, for i=1,2,3. Also assume that $m_1 < m_2$.

$$H_0: \quad F_0(x)+\omega_1 F_1(x)+\omega_2 F_2(x)+ (1-\omega_1-\omega_2)F_3(x) \quad \text{properly.} \quad (2.3.1)$$

Define random variables Y_{1i} and Y_{2i}, for i=1,2,---,n, as follows:

$$Y_{1i} = \begin{cases} 1 & \text{if } X_i < a \\ 0 & \text{otherwise} \end{cases}$$

$$Y_{2i} = \begin{cases} 1 & \text{if } X_i > b \\ 0 & \text{otherwise} \end{cases} ,$$

where a<b are two constants to be defined later.

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\bar{Y}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_{1i}$, $\bar{Y}_2 = \frac{1}{n}\sum_{i=1}^{n} Y_{2i}$.

Under H_0

$$E\bar{X} = \omega_1 m_1 + \omega_2 m_2 + (1-\omega_1-\omega_2)m_3$$

$$E\bar{Y}_1 = \omega_1 P_{11} + \omega_2 P_{12} + (1-\omega_1-\omega_2)P_{13} \quad (2.3.2)$$

$$E\bar{Y}_2 = \omega_1 P_{21} + \omega_2 P_{22} + (1-\omega_1-\omega_2)P_{23}$$

where

$$P_{1i} = \Pr\{X_1 < a \,|\, F_i\} , \quad \text{and}$$

$$P_{2i} = \Pr\{X_i > b \,|\, F_i\} , \quad i=1,2,3. \quad (2.3.3)$$

Eliminating ω_1 and ω_2 in (2.3.2), we obtain

$$E[\bar{X} - \hat{\omega}_1 m_1 - \hat{\omega}_2 m_2 - (1-\hat{\omega}_1-\hat{\omega}_2)m_3] = 0$$

where

$$\hat{\omega}_1 = \frac{(\bar{Y}_1-P_{13})(P_{22}-P_{23}) - (\bar{Y}_2-P_{23})(P_{12}-P_{13})}{(P_{11}-P_{13})(P_{22}-P_{23}) - (P_{21}-P_{23})(P_{12}-P_{13})} \quad (2.3.4)$$

and

$$\hat{\omega}_2 = \frac{(\bar{Y}_1-P_{13})(P_{21}-P_{23})-(\bar{Y}_2-P_{23})(P_{11}-P_{13})}{(P_{12}-P_{13})(P_{21}-P_{23})-(P_{22}-P_{23})(P_{11}-P_{13})} . \qquad (2.3.5)$$

Note that $E\hat{\omega}_1=\omega_1$ and $E\hat{\omega}_2=\omega_2$. Let

$$T_1 = \bar{X} - \hat{\omega}_1 m_1 - \hat{\omega}_2 m_2 - (1-\hat{\omega}_1-\hat{\omega}_2)m_3 , \qquad (2.3.6)$$

then $E(T_1|H_0)=0$.

Proposition 2.3

Under H_0 and the above conditions

$$nVar\ T_1 = \sigma^2 + L_0 + \omega_1 L_1 + \omega_2 L_2 \qquad (2.3.7)$$

where

$$\begin{aligned}
L_0 &= D_1^2 D_3^{-2} P_{13}(1-P_{13}) + D_2^2 D_3^{-2} P_{23}(1-P_{23}) - 2D_1 D_2 D_3^{-2} P_{13} P_{23} \\
&\quad + 2D_1 D_3^{-1} P_{13}(E_{13}-m_3) + 2D_2 D_3^{-1} P_{23}(E_{23}-m_3) \\
L_1 &= -D_1 D_2 D_3^{-2}(P_{11}-P_{13}+P_{21}-P_{23}) + D_1 D_3^{-1}[2P_{11}(E_{11}-m_1) \\
&\quad -2P_{13}(E_{13}-m_3) + (m_1-m_3)(P_{11}+P_{13}-1)] \\
&\quad + D_2 D_3^{-1}[2P_{21}(E_{21}-m_1) - 2P_{23}(E_{23}-m_3) + (m_1-m_3)(P_{21}+P_{23}-1)] \\
L_2 &= -D_1 D_2 D_3^{-2}(P_{22}-P_{23}+P_{12}-P_{13}) + D_1 D_3^{-1}[2P_{12}(E_{12}-m_2) \\
&\quad -2P_{13}(E_{13}-m_3) + (m_2-m_3)(P_{12}+P_{13}-1)] \\
&\quad + D_2 D_3^{-1}[2P_{22}(E_{22}-m_2) - 2P_{23}(E_{23}-m_3) + (m_2-m_3)(P_{12}+P_{13}-1)]
\end{aligned} \qquad (2.3.8)$$

$$\begin{aligned}
D_1 &= (m_1-m_3)(P_{22}-P_{23}) - (m_2-m_3)(P_{21}-P_{23}) \\
D_2 &= (m_2-m_3)(P_{11}-P_{13}) - (m_1-m_3)(P_{12}-P_{13}) \\
D_3 &= (P_{12}-P_{13})(P_{21}-P_{23}) - (P_{22}-P_{23})(P_{11}-P_{13}) \\
E_{1i} &= E(X_1|X_1<a,\ F_i) \\
E_{2i} &= E(X_1|X_1>b,\ F_i), \quad i=1,2,3,
\end{aligned} \qquad (2.3.9)$$

and P_{1i}, P_{2i} are defined by (2.3.3).

Proof: See section 2.6.

Proposition 2.3 shows that nVar T_1 is a linear combination of σ^2+L_0, L_1 and L_2. It was hoped that by assigning appropriate values to a and b, both L_1 and L_2 would vanish. Unfortunately such values could not be found handily. In the following we let

$$a=(m_1+m_3)/2 \quad \text{and} \quad b=(m_2+m_3)/2$$

and show how the expression of nVar T_1 can be simplified partially. It follows from these specific values of a and b that

$$P_{11}+P_{13}=1$$

$$P_{22}+P_{23}=1$$

$$P_{11}(E_{11}-m_1)=P_{13}(E_{13}-m_3)$$

$$P_{22}(E_{22}-m_2)=P_{23}(E_{23}-m_3),$$

and L_1 and L_2 are simplified to the following

$$L_1' = -D_1D_2D_3^{-2}(P_{11}-P_{13}+P_{21}-P_{23})+D_2D_3^{-1}[2P_{21}(E_{21}-m_1)-2P_{23}(F_{23}-m_3) + (m_1-m_3)(P_{21}+P_{23}-1)]$$

$$L_2' = -D_1D_2D_3^{-2}(P_{22}-P_{23}+P_{12}-P_{13})+D_1D_3^{-1}[2P_{12}(E_{12}-m_2)-2P_{13}(E_{13}-m_3) + (m_2-m_3)(P_{12}+P_{13}-1)].$$

Next we will derive a large sample test for H_0. Note first that $\hat{\omega}_1$ and $\hat{\omega}_2$ are both linear functions of $Y_{11},Y_{12},\cdots,Y_{1n},Y_{21},Y_{22},\cdots,Y_{2n}$, and are both means of i.i.d. random variables. Therefore T_1 can be expressed as a mean of i.i.d. random variables. Then by the central limit theorem (*Var* $T_1<\infty$), *under* H_0

$$T_1(\text{Var } T_1)^{-1/2} \to N(0,1) \quad \text{in distribution.}$$

But Var T_1 depends on the unknown parameters ω_1 and ω_2. Using $\hat{\omega}_1$ in (2.3.4) and $\hat{\omega}_2$ in (2.3.5) as estimators of ω_1, ω_2 respectively, we obtain an estimator of Var T_1

$$n\hat{\text{Var}}\ T_1 = \sigma^2 + L_0 + \hat{\omega}L_1' + \hat{\omega}L_2' \qquad (2.3.11)$$

From Var $Y_{11} < \infty$ and Var $Y_{21} < \infty$, it follows that $\hat{\omega}_1 \to \omega_1$ *a.s.* and $\hat{\omega}_1 \to \omega_1$ *a.s.*, hence $\hat{\text{Var}}\ T_1 \to \text{Var}\ T_1$ *a.s.* In the following we state two theorems which will be used to derive large sample results.

Theorem 2.1 (Slutsky, see e.g. Cramer, 1946, pp. 255)

If $X_n, Y_n, \cdots, Z_n$ are random variables converging in probability to the constants $x, y, \cdots, z$ respectively, and rational function $R(X_n, Y_n, \cdots, Z_n)$ converges in probability to the constant $R(x,y,\cdots,z)$, provided that the latter is finite. It follows that any power $R^k(X_n, Y_n, \cdots, Z_n)$ with $k>0$ converges in probability to $R^k(x,y,\cdots,z)$.

Theorem 2.2 (See Cramer, 1946, pp. 254)

Let $X_1, X_2, \cdots$ be a sequence of random variables, with the distribution functions $F_1, F_2, \cdots$. Suppose that $F_n(x)$ tends to a distribution function $F(x)$ as n tends to ∞. Let $Y_1, Y_2, \cdots$ be another sequence of random variables, and suppose that Y_n converges in probability to a constant c. Then the distribution function of $X_n + Y_n$ tends to $F(x-c)$. Further, if $c>0$, the distribution function of $X_n Y_n$ tends to $F(x/c)$, while that of X_n/Y_n tends to $F(cx)$.

By Theorems 2.1 and 2.2 we have

$$T_1(\hat{\text{Var}}\ T_1)^{-1/2} \to N(0,1) \quad \text{in distribution.}$$

Therefore a large sample test can be formulated as

$$\text{Reject } H_0 \quad \text{if} \quad |T_1|(\hat{\text{Var}}\ T_1)^{-1/2} > z_{1-\alpha/2}. \qquad (2.3.12)$$

2.4 Two Components vs. Three Components

Let $X_1,---,X_n$ be a random sample from F_0 and suppose that F_1 and F_2 are each of known form, symmetric about its mean m_i, with common variance σ^2. We wish to test the following hypotheses:

$$H_0 : F_0 = \omega F_1 + (1-\omega)F_2 \text{ properly} \tag{2.4.1}$$

$$H_1 : F_0 = \omega(1-\omega')F_1+(1-\omega)(1-\omega')F_2+\omega'F_3 \text{ properly,} \tag{2.4.2}$$

where ω' is known, $(0<\omega'<1)$, and F_3 has mean m_3, variance σ^2.

Following the argument in section 2.1, we may use $(\hat{\omega}_x-\hat{\omega}_y)\kappa_2^{-1/2}$ as test statistic and reject H_0 if $|\hat{\omega}_x-\hat{\omega}_y|>z_{1-\alpha/2}\cdot\kappa_2^{1/2}$.

Under H_1

$$E(\hat{\omega}_x-\hat{\omega}_y|H_1) = \omega'[(m_1-m_2)^{-1}(m_3-m_2)-(P_1-P_2)^{-1}(P_3-P_2)] \tag{2.4.3}$$

where $P_i = \Pr\{X_1<(m_1+m_2)/2 \mid F_i\}$, i=1,2,3. (2.4.4)

$$\begin{aligned} n\mathrm{Var}(\hat{\omega}_x-\hat{\omega}_y|H_1) = {} & (m_1-m_2)^{-2}\sigma^2 + (P_1-P_2)^{-2}P_1P_2 \\ & -2(m_1-m_2)^{-1}(P_1-P_2)^{-1}[(1-\omega')P_1(E_1-m_1)+\omega'P_3(E_3-m_3)] \\ & +\omega'(1-\omega')[(m_1-m_2)^{-1}(m_3-m_1)-(P_1-P_2)^{-1}(P_3-P_2)] \end{aligned} \tag{2.4.5}$$

where $E_i=E(X_1|X_1<(m_1+m_2)/2, F_i)$, i=1,2,3. Note that $n\mathrm{Var}(\hat{\omega}_x-\hat{\omega}_y|H_1)$ and $E(\hat{\omega}_x-\hat{\omega}_y|H_1)$ depend only on ω', but not on ω. And if $\omega'=0$, then $n\mathrm{Vat}(\hat{\omega}_x-\hat{\omega}_y|H_1) = n\mathrm{Var}(\hat{\omega}_x-\hat{\omega}_y|H_0)$ (the latter is given by (2.1.3)). The asymptotic power of the test then can be calculated by varying F_3 and ω in the following formula:

$$\begin{aligned} & \Pr\{|\hat{\omega}_x-\hat{\omega}_y|>z_{1-\alpha/2}\cdot\kappa_2^{-1/2} \mid H_1\} \\ & =\Phi([\mathrm{Var}(\hat{\omega}_x-\hat{\omega}_y|H_1)]^{-1/2}[-z_{1-\alpha/2}\kappa_2^{-1/2} - E(\hat{\omega}_x-\hat{\omega}_y|H_1)]) \\ & +\Phi([(\mathrm{Var}(\hat{\omega}_x-\hat{\omega}_y|H_1)]^{-1/2}[-z_{1-\alpha/2}\kappa_2^{-1/2} + E(\hat{\omega}_x-\hat{\omega}_y|H_1)]) \end{aligned} \tag{2.4.6}$$

2.5 Proofs of Propositions 2.1 and 2.2

$\{X_1, X_2, \cdots, X_n\}$ is a random sample from F_0. F_1 and F_2 are under the same conditions as in section 2.1. For convenience suppose that X and X_1 have the same distribution F_0, and Y and Y_1 have the same distribution. By definition

$$\hat{\omega}_x = (\bar{X}-m_2)(m_1-m_2)^{-1} \quad \text{and} \quad \hat{\omega}_y = (\bar{Y}-P_2)(P_1-P_2)^{-1} .$$

Let $U=X-EX$ and $V=Y-EY$. Then we have

$$EX = \omega m_1 + (1-\omega)m_2$$

$$EY = \omega P_1 + (1-\omega)P_2$$

$$EU = EV = 0$$

$$EU^2 = \text{Var } X = \sigma^2 + \omega(1-\omega)(m_1-m_2)^2$$

$$EV^2 = \text{Var } Y = P_1P_2 + \omega(1-\omega)(P_1-P_2)^2$$

$$EUV = \text{Cov}(X,Y) = P_1(E_1-m_1) + \omega(1-\omega)(m_1-m_2)(P_1-P_2)$$

$$EU^3 = \omega(1-\omega)(1-2\omega)(m_1-m_2)^3$$

$$EU^2V = (1-2\omega)[-\int_{-\infty}^{\theta}(x-m_1)^2dF_1(x)+P_1\sigma^2]+\omega(1-\omega)(1-2\omega)(m_1-m_2)^2(P_1-P_2)$$

where $\theta=(m_1+m_2)/2$,

$$EUV^2 = (1-2\omega)(P_1-P_2)[P_1(E_1-m_1)+\omega(1-\omega)(m_1-m_2)(P_1-P_2)]$$

$$EV^3 = [P_1P_2+\omega(1-\omega)(P_1-P_2)^2](1-2\omega)(P_1-P_2)$$

$$EU^4 = \int_{-\infty}^{\infty}(x-m_1)^4dF_1(x)+6\omega(1-\omega)(m_1-m_2)^2\sigma^2+\omega(1-\omega)(1-3\omega+3\omega^2)(m_1-m_2)^4$$

$$EU^3V = \int_{-\infty}^{\theta}(x-m_1)^3dF_1(x)+3\omega(1-\omega)(m_1-m_2)[2\int_{-\infty}^{\theta}(x-m_1)^2dF_1(x)-\sigma^2]$$
$$+3\omega(1-\omega)(m_1-m_2)^2P_1(E_1-m_1)+\omega(1-\omega)(1-3\omega+3\omega^2)(m_1-m_2)^3(P_1-P_2)$$

$$EU^2V^2 = -(1-2\omega)^2(P_1-P_2)\int_{-\infty}^{\theta}(x-m_1)^2dF_1(x)+(1-2\omega)^2(P_1-P_2)P_1\sigma^2$$
$$+\omega(1-\omega)(1-3\omega+3\omega^2)(m_1-m_2)^2(P_1-P_2)^2+[P_1P_2+\omega(1-\omega)(P_1-P_2)^2]\sigma^2$$
$$+\omega(1-\omega)(m_1-m_2)^2P_1P_2$$

$$EUV^3 = [P_1P_2+(1-3\omega+3\omega^2)(P_1-P_2)^2][P_1(E_1-m_1)+\omega(1-\omega)(m_1-m_2)(P_1-P_2)]$$

$$EV^4 = P_1^2P_2^2 + (1-2\omega+2\omega^2)P_1P_2(P_1-P_2)^2+\omega(1-\omega)(1-3\omega+3\omega^2)(P_1-P_2)^4$$

Proof of Proposition 2.1:

$$n^2E(\hat{\omega}_x-\hat{\omega}_y)^3 = (m_1-m_2)^{-3}EU^3 -3(m_1-m_2)^{-2}(P_1-P_2)^{-1}EU^2V$$
$$+3(m_1-m_2)^{-1}(P_1-P_2)^{-2}EUV^2 - (P_1-P_2)^{-3}EV^3$$

$$=(1-2\omega)\{3(m_1-m_2)^{-2}(P_1-P_2)^{-1}[\int_{-\infty}^{\theta}(x-m_1)^2dF_1(x)-P_1\sigma^2]$$
$$+3P_1(E_1-m_1)(m_1-m_2)^{-1}(P_1-P_2)^{-1}-P_1P_2(P_1-P_2)^{-2}\}$$

Proof of Proposition 2.2:

$$n^3\kappa_4 = (m_1-m_2)^{-4}[EU^4-3(EU^2)^2] - 4(m_1-m_2)^{-3}(P_1-P_2)^{-1}(-3EU^2EUV+EU^3V)$$
$$+6(m_1-m_2)^{-2}(P_1-P_2)^{-2}[EU^2V^2-EU^2\cdot EV^2-2(EUV)^2]$$
$$-4(m_1-m_2)^{-1}(P_1-P_2)^{-3}(EUV^3-3EUV\cdot EV^2)+(P_1-P_2)^{-4}[EV^4-3(EV^2)^2]$$

$$= (m_1-m_2)^{-4}[\int_{-\infty}^{\infty}(x-m_1)^4dF_1(x)-3\sigma^4]-4(m_1-m_2)^{-3}(P_1-P_2)^{-1}\cdot$$
$$[\int_{-\infty}^{\theta}(x-m_1)^3dF_1(x)-3\sigma^2P_1(E_1-m_1)]$$
$$-6(m_1-m_2)^{-2}(P_1-P_2)^{-1}[\int_{-\infty}^{\theta}(x-m_1)^2dF_1(x)- P_1\sigma^2]$$
$$-12(m_1-m_2)^{-2}(P_1-P_2)^{-2}P_1^2(E_1-m_1)^2$$
$$-4(m_1-m_2)^{-1}(P_1-P_2)^{-3}\ (1-6P_1P_2)P_1(E_1-m_1)$$
$$+(P_1-P_2)^{-4}\ P_1P_2(1-6P_1P_2)$$

2.6 Proof of proposition 2.3

We use the same notation as in section 2.3. By definition

$$T_1 = \bar{X} -\hat{\omega}_1m_1 -\hat{\omega}_2m_2 - (1-\hat{\omega}_1 -\hat{\omega}_2)m_3$$

$$= \bar{X}-m_3 + D_1D_3^{-1}(\bar{Y}_1-P_{13}) + D_2D_3^{-1}(\bar{Y}_2-P_{23}),$$

where D_1, D_2 and D_3 are defined as in Proposition 2.3.

By definition we have

$$\text{Var } X_1 = \sigma^2+\omega_1(1-\omega_1)(m_1-m_3)^2+\omega_2(1-\omega_2)(m_2-m_3)^2-2\omega_1\omega_2(m_1-m_3)(m_2-m_3)$$

$$\begin{aligned}\text{Var } Y_{11} = {} & P_{13}(1-P_{13})+\omega_1(P_{11}-P_{13})(1-2P_{13})+\omega_2(P_{12}-P_{13})(1-2P_{13}) \\ & -\omega_1^2(P_{11}-P_{13})^2-\omega_2^2(P_{12}-P_{13})^2-2\omega_1\omega_2(P_{11}-P_{13})(P_{12}-P_{13})\end{aligned}$$

$$\begin{aligned}\text{Var } Y_{21} = {} & P_{23}(1-P_{23})+\omega_1(P_{21}-P_{23})(1-2P_{23})+\omega_2(P_{22}-P_{23})(1-2P_{23}) \\ & -\omega_1^2(P_{21}-P_{23})^2-\omega_2^2(P_{22}-P_{23})^2-2\omega_1\omega_2(P_{21}-P_{23})(P_{22}-P_{23})\end{aligned}$$

$$\begin{aligned}\text{Cov}(X_1,Y_{11}) = {} & P_{13}(E_{13}-m_3)+\omega_1[P_{11}(E_{11}-m_1)-P_{13}(E_{13}-m_3) \\ & +(m_1-m_3)(P_{11}-P_{13})]-\omega_1^2(m_1-m_3)(P_{11}-P_{13})-\omega_2^2(m_2-m_3)(P_{12}-P_{13}) \\ & +\omega_2[P_{12}(E_{12}-m_2)-P_{13}(E_{13}-m_3)+(m_2-m_3)(P_{12}-P_{13})] \\ & -\omega_1\omega_2[(m_1-m_3)(P_{12}-P_{13})+(m_2-m_3)(P_{11}-P_{13})]\end{aligned}$$

$$\begin{aligned}\text{Cov}(X_1,Y_{21}) = {} & P_{23}(E_{23}-m_3)-\omega_1^2(m_1-m_3)(P_{21}-P_{23})-\omega_2^2(m_2-m_3)(P_{22}-P_{23}) \\ & +\omega_1[P_{21}(E_{21}-m_1)-P_{23}(E_{23}-m_3)+(m_1-m_3)(P_{21}-P_{23})] \\ & +\omega_2[P_{22}(E_{22}-m_2)-P_{23}(E_{23}-m_3)+(m_2-m_3)(P_{22}-P_{23})] \\ & -\omega_1\omega_2[(m_1-m_3)(P_{22}-P_{23})+(m_2-m_3)(P_{21}-P_{23})]\end{aligned}$$

$$\begin{aligned}\text{Cov}(Y_{11},Y_{21}) = {} & P_{13}P_{23}+\omega_1^2(P_{11}-P_{13})(P_{21}-P_{23})+\omega_2^2(P_{12}-P_{13})(P_{22}-P_{23}) \\ & +\omega_1[P_{23}(P_{11}-P_{13})+P_{13}(P_{21}-P_{23})]+\omega_2[P_{13}(P_{22}-P_{23})+P_{23}(P_{12}-P_{13})] \\ & +\omega_1\omega_2[(P_{11}-P_{13})(P_{22}-P_{23})+(P_{12}-P_{13})(P_{21}-P_{23})]\end{aligned}$$

Therefore

$$nD_3^2 \text{Var } T_1 = D_3^2 \text{Var } X_1 + D_1^2 \text{Var } Y_{11} + D_2^2 \text{Var } Y_{21} + 2D_1D_3 \text{Cov}(X_1, Y_{11})$$

$$+2D_2D_3 \text{Cov}(X_1, Y_{21}) + 2D_1D_2 \text{Cov}(Y_{11}, Y_{21})$$

can be expressed as a linear combination of ω_1, ω_2, ω_1^2, ω_1^2 and $\omega_1\omega_2$.

The coefficient of ω_1 is

$$D_3^2(m_1-m_3)^2+D_1^2(P_{11}-P_{13})(1-2P_{13})+D_2^2(P_{21}-P_{23})(1-2P_{23})$$

$$-2D_1D_2[P_{23}(P_{11}-P_{13})+P_{13}(P_{21}-P_{23})]+2D_1D_3[P_{11}(E_{11}-m_1)-P_{13}(E_{13}-m_3)$$

$$+(m_1-m_3)(P_{11}-P_{13})]+2D_2D_3[P_{21}(E_{21}-m_1)-P_{23}(E_{23}-m_3)+(m_1-m_3)(P_{21}-P_{23})]$$

$$= D_3^2L_1 .$$

The coefficient of ω_2 is

$$D_3^2(m_2-m_3)^2+D_1^2(P_{12}-P_{13})(1-2P_{13})+D_2^2(P_{22}-P_{23})(1-2P_{23})$$

$$-2D_1D_2[P_{13}(P_{22}-P_{23})+P_{23}(P_{12}-P_{13})]+2D_1D_3[P_{12}(E_{12}-m_2)-P_{13}(E_{13}-m_3)$$

$$+(m_2-m_3)(P_{12}-P_{13})]+2D_2D_3[P_{22}(E_{22}-m_2)-P_{23}(E_{23}-m_3)+(m_2-m_3)(P_{22}-P_{23})]$$

$$= D_3^2L_2 .$$

The coefficient of ω_1^2 is

$$-[D_3(m_1-m_3)+D_1(P_{11}-P_{13})+D_2(P_{21}-P_{23})]^2 = 0 .$$

The coefficient of ω_2^2 is

$$-[D_3(m_2-m_3)+D_1(P_{12}-P_{13})+D_2(P_{22}-P_{23})]^2 = 0 .$$

The coefficient of $\omega_1\omega_2$ is

$$-[D_3(m_1-m_3)+D_1(P_{11}-P_{13})+D_2(P_{21}-P_{23})][D_3(m_2-m_3)+D_1(P_{12}-P_{13})$$

$$+D_2(P_{22}-P_{23})] = 0 .$$

The constant term is

$$D_3^2\sigma^2+D_1^2P_{13}(1-P_{13})+D_2^2P_{23}(1-P_{23})-2D_1D_2P_{13}P_{23}-2D_1D_3P_{13}(E_{13}-m_3)$$

$$+2D_2D_3P_{23}(E_{23}-m_3) = D_3^2(\sigma^2+L_0) .$$

CHAPTER III

TESTS USING RANK-TYPE STATISTICS

In this chapter we describe a rank-type statistic proposed by Thomas(1969), in the form used by Hariton(1972), and modify it in order to apply to problems A2 and A3. Next we compare the performance of this modified statistic with Johnson's statistics in the case of two normal components against a single normal alternative. Then we extend the method to obtain test statistics for three and four components mixtures, and give an algorithm to derive test statistics for mixtures of more than four components. Finally in section 3.8 we study two special cases of reducing the number of components of a given finite proper mixtures.

3.1 Problem A1

Suppose that $X_{i1}, X_{12}, \cdots, X_{in_i}$ is a random sample from the c.d.f. F_i, i=0,1,2 and that all the X's are mutually independent. We wish to test the following null hypothesis

$H_0 : F_0 = \omega F_1 + (1 - \omega)F_2$ properly.

Thomas(1969), assuming (i) the continuity of F_0, F_1 and F_2, and (ii) $n_0=n_1=n_2$, proposed a statistic to test H_0 and showed its asymptotic normality. Hariton(1972), assuming only condition (i), expresses this statistic as

$$T_2 = W_{10} + W_{02} - W_{12} - 1/2 \qquad (3.1.1)$$

where

$$W_{ij} = \frac{1}{n_i n_j} \sum_{t=1}^{n_i} \sum_{k=1}^{n_j} h(X_{jk}-X_{it}), \tag{3.1.2}$$

and $h(x)= \begin{cases} 1 & \text{if } x>0 \\ 0 & \text{otherwise.} \end{cases}$

Define

$$\alpha_{ij}= \int F_i dF_j = Pr\{X_{i1}<X_{j1}\}, \tag{3.1.3}$$

then

(1) $\alpha_{ii}=1/2$, $\alpha_{ji}=1-\alpha_{ij}$ (by continuity)

(2) $EW_{ij} = \alpha_{ij}$

(3) $$\text{Var } W_{ij} = \frac{1}{n_i n_j}[\alpha_{ij}-(n_i+n_j-1)\alpha_{ij}^2+(n_i-1)\int F_i^2 dF_j + (n_j-1)\int (1-F_j)^2 dF_i] \tag{3.1.4}$$

(4) $$\text{Cov}(W_{ij},W_{ik}) = \frac{1}{n_i}[\int F_j F_k dF_i - \alpha_{ji}\alpha_{ki}]$$

(For proofs of (2),(3) and (4), see Birnbaum and Klose(1957).)

Integrating both sides of the equation $F_0=\omega F_1+(1-\omega)F_2$ with respect to F_0, F_1, F_2, we have

$$1/2 = \omega\alpha_{10}+(1-\omega)\alpha_{20}$$

$$\alpha_{01} = \omega/2 + (1-\omega)\alpha_{21}$$

$$\alpha_{02} = \omega\alpha_{12}+(1-\omega)/2 .$$

Solving for ω in each equation and equating these solutions by pairs, we obtain three conditions on α's. Hariton shows that these three conditions are mutually equivalent. Hence under H_0 a (single) necessary condition is therefore

$$\alpha_{10}+\alpha_{02}-\alpha_{12}-1/2=0. \tag{3.1.5}$$

It follows that under H_0 $ET_2=0$.

Let $N=n_0+n_1+n_2$ and suppose that as $N \to \infty$

$n_i/N \to r_i$, i=0,1,2 and there exists ε such that $0<\varepsilon\le r_i\le 1-\varepsilon<1$. Then by the following Theorems 3.1 and 3.2, T_2 has an asymptotic normal distribution.

Theorem 3.1

Let $F_1, F_2, ---, F_k$ be k univariate c.d.f.'s and suppose that for i=1,2,---,k, a random sample $\{X_{i1}, ---, X_{in_i}\}$ of size n_i is available from F_i. Assume that all X's are mutually independent. Let $N=n_1+---+n_k$, $r_i=n_i/N$, i=1,---,k, and also assume that there exists ε such that $0<\varepsilon\le r_i\le 1-\varepsilon<1$, i=1,---,k. Then for any integer m, the random vector

$$N^{1/2}((W_{i_1,j_1}-1/2),(W_{i_2,j_2}-1/2),---,(W_{i_m,j_m}-1/2))$$

has an asymptotic multi-normal distribution with finite mean vector and finite variance-covariance matrix provided i_t's and j_t's are integers such that $1\le i_t\ne j_t\le k$.

Proof: As stated in Hariton(1972) this follows from Theorem 5.5.1 of Puri and Sen (1971), pp. 196 by setting c=2 and

$$J_{N(i)}\left(\frac{\alpha}{N+1}\right) = E^{(i)}_{N,\alpha} = \frac{\alpha}{N+1}, \quad 1\le\alpha\le N.$$

Theorem 3.2

Let X be a k×1 vector of random variables, and let $\{X^{(1)}, --- X^{(n)}, ---\}$ be a sequence of vectors of random variables such that $X^{(n)} \to X$ in distribution. Let $g_1, ---, g_m$ be continuous functions defined on R^k. Then

$(g_1(X^{(n)}), ---g_m(X^{(n)})) \to (g_1(X), ---g_m(X))$ in distribution.

Proof: See Breiman (1968), pp. 237.

$VarT_2$ can be derived by using the formulae in (3.1.4). But since terms in (3.1.4) involve unknown distributions F_0, F_1, F_2, these terms

can only be estimated. One possible way is to use the following extended theorems(see Hariton, 1972) of Woinsky and Kurz (1969, pp. 447):

$$\frac{1}{n_a n_b n_c} \sum_{i=1}^{n_a} \sum_{j=1}^{n_b} \sum_{k=1}^{n_c} h(x_{ck}-X_{ai})h(X_{ck}-X_{bj}) \to \int F_b F_a dF_c \text{ in probability,}$$

$$2W_{ab}-1+\frac{1}{n_b^2 n_a} \sum_{i=1}^{n_b} \sum_{j=1}^{n_b} \sum_{k=1}^{n_a} h(X_{ak}-X_{bi})h(X_{ak}-X_{bj}) \to \int (1-F_b)^2 dF_a$$

in probability. (3.1.6)

Denote the (consistent) estimator of Var T_2 obtained in this way by $\hat{\text{Var}}\ T_2$. Hariton then obtains the following large sample test

$$\text{Reject } H_0 \text{ if } \quad |T_2|(\hat{\text{Var}}\ T_2)^{-1/2} > z_{1-\alpha/2}\ , \tag{3.1.7}$$

where $z_{1-\alpha/2}$ is such that $\Phi(z_{1-\alpha/2})=1-\alpha/2$.

3.2 Problem A2

Suppose now that F_2 is known and $X_{i1},\cdots X_{in_i}$ are random samples from F_i, i=0,1 with all the X's mutually independent. Assume that F_0, F_1, and F_2 are continuous. In order to test the following null hypothesis

$$H_0 : F_0=\omega F_1 +(1-\omega)F_2 \text{ properly,}$$

we define random variables

$$R_{ij} = \int F_{in_i}(x)\ dF_j(x)$$

$$=1-\frac{1}{n_i}\sum_{k=1}^{n_i} F_j(X_{ik}) \tag{3.2.1}$$

where F_{in_i} is the e.d.f. corresponding to $X_{i1},\cdots,X_{in_i}$. Note that $ER_{ij}=\alpha_{ij}$. Define a statistic

$$T_3 = W_{10} + R_{02} - R_{12} - 1/2 \tag{3.2.2}$$

where W_{10} is defined by (3.1.2).

Under H_0,

$$ET_3 = \alpha_{10} + \alpha_{02} - \alpha_{12} - \frac{1}{2} = 0 .$$

$$\text{Var } R_{i2} = ER_{i2}^2 - \alpha_{i2}^2$$

$$= \frac{1}{n_i}\left[-\alpha_{2i}^2 + \int F_2^2(x)dF_i(x)\right] , \quad \text{for } i=0,1. \qquad (3.2.3)$$

$$\text{Cov}(R_{02},R_{12}) = 0$$

$$\text{Cov}(W_{10},R_{12}) = E\left\{\left[\frac{1}{n_0n_1}\sum_{i=1}^{n_1}\sum_{j=1}^{n_0} h(X_{0j}-X_{1i})\right]\left[1-\frac{1}{n_1}\sum_{k=1}^{n_1} F_2(X_{1k})\right]\right\} - \alpha_{10}\alpha_{12}$$

$$= \alpha_{10} - \frac{1}{n_0n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_0}\sum_{k=1}^{n_1} Eh(X_{0j}-X_{1i})F_2(X_{1k}) - \alpha_{10}\alpha_{12}$$

$$= \alpha_{10} - \alpha_{10}\alpha_{12} - \frac{1}{n_0n_1^2}\left[n_0n_1\int (1-F_0(x))F_2(x)dF_1(x) + n_1n_0(n_1-1)\alpha_{10}\alpha_{21}\right]$$

$$= \frac{1}{n_1}\left[-\alpha_{01}\alpha_{21} + \int F_0(x)F_2(x)dF_1(x)\right]$$

$$\text{Cov}(W_{10},R_{02}) = \frac{1}{n_0}\left[\alpha_{10}\alpha_{20} - \int F_1(x)F_2(x)dF_0(x)\right] , \qquad (3.2.4)$$

and $\text{Var}W_{10}$ follows from (3.1.4).

Var T_3 can then be calculated as a linear combination of the above terms. As in section 3.1, it is desirable to have a consistent estimator for Var T_3, so that large sample tests can be applied. For this purpose, we need the following:

Proposition 3.1

(1) $R_{i2} \xrightarrow{p} \alpha_{i2}$

(2) $\frac{1}{n_0n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} h(X_{1j}-X_{0i})F_2(X_{1j}) \xrightarrow{p} \int F_0(x)F_2(x)dF_1(x)$

$$(3)\ \frac{1}{n_a}\sum_{i=1}^{n_a}[F_2(X_{ai})]^2 \xrightarrow{p} \int F_2^2(x)dF_a(x),\quad a=0,1$$

$$(4)\ \frac{1}{n_1^2 n_0}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\sum_{k=1}^{n_0} h(X_{0k}-X_{1i})h(X_{0k}-X_{1j}) \xrightarrow{p} \int F_1^2(x)dF_0(x)$$

as $n_0, n_1, n_a \longrightarrow \infty$.

Proof: (4) follows from (3.1.6),

Since $ER_{i2} = \alpha_{i2}$,

$Eh(X_{1j}-X_{0i})F_2(X_{1j}) = \int F_0(x)F_2(x)dF_1(x)$ for $j=1,\cdots, n_1$; $i=1,\cdots,n_0$.

$E[F_2(X_{ai})]^2 = \int F_2^2(x)dF_a(x)$, $a=0,1$.

(1) and (3) follow from the weak law of large numbers. (Since R_{0i} is a mean of i.i.d. random variables). Under the assumption of continuity, after rearranging the terms,

$\frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} h(X_{1j}-X_{0i})F_2(X_{1j}) = \int F_{0n_0}(x)F_2(x)dF_{1n_1}(x)$. Since $F_{an_a}(x) \longrightarrow F_a(x)$ *a.s.* (by the strong law of large numbers, and continuity of $F_a(x)$), $\left|\int F_{0n_0}(x)F_2(x)dF_{1n_1}(x) - \int F_0(x)F_2(x)dF_1(x)\right| \le$

$$\left|\int (F_{0n_0}(x) - F_0(x)F_2(x)dF_{1n_1}(x)\right| + \left|\int F_0(x)F_2(x)dF_{1n_1}(x) - \int F_0(x)F_2(x)dF_1(x)\right|$$

$$= \left|\frac{1}{n_1}\sum_{i=1}^{n_1}[F_{0n_0}(X_{1i})-F_0(X_{1i})]F_2(X_{1i})\right|$$

$$+ \left|\frac{1}{n_1}\sum_{i=1}^{n_1} F_0(X_{1i})F_2(X_{1i}) - \int_{-\infty}^{\infty} F_0(x)F_2(x)dF_1(x)\right|$$

$\longrightarrow 0$ *a.s.* (by the strong law of large numbers) q.e.d.

After rearrangement,

$$\text{Var } T_3 = \frac{1}{n_0}\int [F_1(x)-F_2(x)]^2 dF_0(x) + \frac{1}{n_1}\int [F_0(x)-F_2(x)]^2 dF_1(x)$$

$$- \frac{1}{n_0 n_1}\left[\int F_1^2(x)dF_0(x) + \int F_0{}^2(x)dF_1(x)\right] - \frac{1}{n_0}(\alpha_{10} - \alpha_{20})^2$$

$$- \frac{1}{n_1}(\alpha_{01} - \alpha_{21})^2 + \frac{1}{n_0 n_1}(\alpha_{10}{}^2 - \alpha_{10} + 1)$$

Therefore a consistent estimator for Var T_3 is

$$\hat{\text{Var}}\ T_3 = \frac{n_1 - 1}{n_0 n_1}\left[\frac{1}{n_1{}^2 n_0} \sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\sum_{k=1}^{n_0} h(X_{0k} - X_{1i})h(X_{0k} - X_{1j})\right]$$

$$+ \frac{n_0 - 1}{n_0 n_1}\left[\frac{1}{n_0{}^2 n_1} \sum_{i=1}^{n_0}\sum_{j=1}^{n_0}\sum_{k=1}^{n_1} h(X_{1k} - X_{0i})h(X_{1k} - X_{0j})\right]$$

$$+ \frac{1}{n_0}\left[\frac{1}{n_0} \sum_{i=1}^{n_0} F_2{}^2(X_{0i})\right] + \frac{1}{n_1}\left[\frac{1}{n_1} \sum_{i=1}^{n_1} F_2{}^2(X_{1i})\right]$$

$$- \frac{2}{n_0}\left[\frac{1}{n_0 n_1} \sum_{i=1}^{n_1}\sum_{j=1}^{n_0} h(X_{0j} - X_{1i})F_2(X_{0j})\right]$$

$$- \frac{2}{n_1}\left[\frac{1}{n_0 n_1} \sum_{i=1}^{n_0}\sum_{j=1}^{n_1} h(X_{1j} - X_{0i})F_2(X_{1j})\right]$$

$$- \frac{1}{n_0}(W_{10} - R_{20})^2 - \frac{1}{n_1}(W_{01} - R_{21})^2 + \frac{1}{n_0 n_1}(W_{10}{}^2 - W_{10} + 1) \qquad (3.2.5)$$

Let $N = n_0 + n_1$, and suppose that as $N \to \infty$, $n_i/N \to \gamma_i$, $i=0,1$, and there exists ε such that $0 < \varepsilon \le \gamma_i \le 1-\varepsilon < 1$. Then by Theorems 2.1, 2.2, 3.1, and 3.2, under H_0

$T_3(\hat{\text{Var}}\ T_3)^{-1/2} \longrightarrow N(0,1)$ in distribution. Therefore a large sample test can be formulated as follows:

Reject H_0 if $|T_3|(\hat{\text{Var}}\ T_3)^{-1/2} > z_{1-\alpha/2}$, (3.2.6)

where $z_{1-\alpha/2}$ is such that $\Phi(z_{1-\alpha/2}) = 1 - \alpha/2$.

3.3 Problem A3

Suppose that F_1 and F_2 are known, and that $X_1, X_2, \cdots, X_n$ is a random sample from F_0. Also assume that F_0, F_1 and F_2 are continuous. The null hypothesis is

H_0: $F_0(x) = \omega F_1(x) + (1-\omega) F_2(x)$ properly

First define random variables

$R_{0i} = \int F_n(x) dF_i(x)$, $i=1,2$, where $F_n(x)$ is the empirical distribution function corresponding to $X_1, X_2, \cdots, X_n$.

Define the test statistic

$$T_4 = R_{10} + R_{02} - \alpha_{12} - \frac{1}{2}. \qquad (3.3.1)$$

Then following an argument similar to that in section 3.3, when H_0 is true, we have

$$ET_4 = \alpha_{10} + \alpha_{02} - \alpha_{12} - \frac{1}{2}.$$

Var R_{10} and Var R_{02} can be similarly derived as in (3.2.3) by changing indices. We have

$$\begin{aligned} \mathrm{Cov}(R_{10}, R_{02}) &= -\mathrm{Cov}(R_{01}, R_{02}) \\ &= -E\left\{[1 - \frac{1}{n}\sum_{i=1}^{n} F_1(X_i)][1 - \frac{1}{n}\sum_{i=1}^{n} F_2(X_i)]\right\} - \alpha_{01}\alpha_{02} \\ &= -1 + \alpha_{10} + \alpha_{20} - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E[F_1(X_i)F_2(X_j)] + \alpha_{01}\alpha_{02} \\ &= \frac{1}{n}[\alpha_{10}\alpha_{20} - \int F_1(x)F_2(x)dF_0(x)] \qquad (3.3.2) \end{aligned}$$

It follows that

$$\begin{aligned} \mathrm{Var}\, T_4 &= \mathrm{Var}\, R_{10} + \mathrm{Var}\, R_{02} + 2\mathrm{Cov}(R_{10}, R_{02}) \\ &= \frac{1}{n}\left\{-(\alpha_{10} - \alpha_{20})^2 + \int [F_1(x) - F_2(x)]^2 dF_0(x)\right\} \qquad (3.3.3) \end{aligned}$$

Hence a consistent estimator for Var T_4 is

$$\hat{\text{Var}}\ T_4 = \frac{1}{n}\left\{-(R_{10} - R_{20})^2 + \frac{1}{n}\sum_{i=1}^{n}[F_1(X_i)-F_2(X_i)]^2\right\} \qquad (3.3.4)$$

By the central limit theorem and Theorems 2.1, 2.2, 3.1 and 3.2, $T_4(\text{Var}\ T_4)^{-1/2}$ and $T_4(\hat{\text{Var}}\ T_4)^{-1/2}$ have a same asymptotic distribution, viz., a univariate normal distribution N(0,1). Thus a large sample test can be formulated as follows:

Reject H_0 if $|T_4| > z_{1-\alpha/2}(\hat{\text{Var}}\ T_4)^{1/2}$

where $z_{1-\alpha/2}$ is such that $\Phi(z_{1-\alpha/2}) = 1-\alpha/2$.

3.4 Two Normal Components vs. a Single Normal Alternative

In this section we study the performance of T_4 with respect to the same hypotheses, H_0, H_1, as in section 2.2 and then compare its approximate power with other tests. Assume that $F_i \sim N(m_i,\sigma^2)$, $i=1,2$. The hypotheses to be considered are

H_0: $F_0 = \omega F_1 + (1-\omega)F_2$ properly

H_1: $F_0 \sim N(\mu,\tau^2)$.

By definition

$$T_4 = R_{10} + R_{02} - \alpha_{12} - \frac{1}{2}$$

where R_{0i} is defined by (3.2.1) and α_{12} by (3.1.3). It follows that

$$ET_4 = \alpha_{10} + \alpha_{02} - \alpha_{12} - \frac{1}{2} \qquad \text{and}$$

$$n\text{Var}\ T_4 = -(\alpha_{10} - \alpha_{20})^2 + \int[F_1(x) - F_2(x)]^2 dF_0(x) .$$

Under H_0

$$E(T_4|H_0) = 0$$

$$n\text{Var}(T_4|H_0) = -(\tfrac{1}{2} - \alpha_{21})^2 + \int[F_1(x) - F_2(x)]^2 dF_2(x)$$

$$+\omega\int[F_1(x) - F_2(x)]^2 d[F_1(x) - F_2(x)] \qquad (3.4.1)$$

If, furthermore, we let $m_2=-m_1$, then the last term on the right hand side of (3.4.1) vanishes. Therefore

$$nVar(T_4|H_0) = -(\frac{1}{2} - \alpha_{21})^2 + \int[F_1(x) - F_2(x)]^2 dF_2(x)$$

$$= -[\frac{1}{2} - \Phi(\sqrt{2}m_1\sigma^{-1})]^2 + \int_{-\infty}^{\infty}[\Phi(\frac{x-m_1}{\sigma}) - \Phi(\frac{x+m_1}{\sigma})]^2\phi(\frac{x+m_1}{\sigma})dx \quad (3.4.2)$$

where $\phi(x)$ is the derivative of $\Phi(x)$ and $\Phi(x)$ is the cumulative standard normal distribution. Note that in (3.4.2) $nVar(T_4\ H_0)$ does not depend on ω. A large sample test can be formulated as follows:

$$\text{Reject } H_0 \text{ if } |T_4|[Var(T_4|H_0)]^{-1/2} > z_{1-\alpha/2} \quad (3.4.3)$$

where $z_{1-\alpha/2}$ is such that $\Phi(z_{1-\alpha/2}) = 1-\alpha/2$ and α is the asymptotic significance level.

Under H_1

$$E(T_4|H_1) = \Phi(\frac{\mu-m_1}{\sqrt{\sigma^2+\tau^2}}) + \Phi(\frac{-\mu-m_1}{\sqrt{\sigma^2+\tau^2}}) - \Phi(-\sqrt{2}m_1\sigma^{-1}) - \frac{1}{2} \quad \text{and}$$

$$nVar(T_4|H_1) = -[\Phi(\frac{\mu-m_1}{\sqrt{\sigma^2+\tau^2}}) - \Phi(\frac{\mu+m_1}{\sqrt{\sigma^2+\tau^2}})]^2 + \int_{-\infty}^{\infty}[\Phi(\frac{x-m_1}{\sigma}) - \Phi(\frac{x+m_1}{\sigma})]^2\phi(\frac{x-\mu}{\tau})dx \quad (3.4.4)$$

For a function $f(z)$ having no pole between the real axis and the lines $z=\pm i\pi h$, it can be shown by contour integration that (Goodwin, 1949):

$$\int_{-\infty}^{\infty} f(x)e^{-x^2}dx = h\sum_{n=-\infty}^{\infty} f(nh)e^{-n^2h^2} + R(h) \quad (3.4.5)$$

where $|R(h)| \leq 2\sqrt{\pi}e^{-\pi^2/h^2}$.

We first use (3.4.5) with h=1 to calculate the integrals in (3.4.2) and (3.4.3) for various values of m_1, μ, τ with $\sigma^2=1$. When h=1

$$|R(1)| \leq 2\sqrt{\pi}\ e^{-\pi^2} \simeq .00018\ ,$$

i. e. the approximation (3.4.5) is accurate to 3 decimals with maximal error ±.00018. From these values we then obtain the approximate power of the test (3.4.3) with respect to the alternative hypothesis H_0 from the formula

$$\Phi\left(\frac{-z_{1-\alpha/2}[\mathrm{Var}(T_4|H_0)]^{1/2} + E(T_4|H_0)}{[\mathrm{Var}(T_4|H_1)]^{1/2}}\right) + \Phi\left(\frac{-z_{1-\alpha/2}[\mathrm{Var}(T_4|H_0)]^{1/2} - E(T_4|H_0)}{[\mathrm{Var}(T_4|H_1)]^{1/2}}\right) \tag{3.4.6}$$

Table 3.1 contains the approximate power of the test (3.4.3) for 5% significance level with $\sigma^2=1$, $m_2=-m_1$, and various values of m_1, μ, and τ. Comparing Table 3.1 with Table 2.3, we see that for the case of two normal components versus a single normal alternative, the test using T_4 (i.e. (3.4.3)) is more powerful than the test using $\hat{\omega}_x - \hat{\omega}_y$ (i. e. (2.1.3)), especially when $\tau=1,2$. Next, comparing Table 3.1, for n=100,400, with Johnson's table IIb (1973, pp.24) which uses another statistic, we see that in most places the power is fairly comparable except that when $\tau=2$, $m_1=1$, n=100, the test using T_4 is considerably more powerful.

3.5 Problems A1', A2', A3' —— Three Components

In this section, the following null hypothesis will be tested:

$$H_0: \; F_0(x) = \omega_1 F_1(x) + \omega_2 F_2(x) + (1-\omega_1-\omega_2)F_3(x) \quad \text{properly.} \tag{3.5.1}$$

Let $X_{01}, \text{---}, X_{0n_0}$ be a random sample from F_0 and suppose that each of F_1, F_2, F_3, is either known or there is a random sample from it. Also assume that all the r.v.'s in the samples are mutually independent and F_0, F_1, F_2, F_3 are continuous.

Integrating both sides of (3.5.1) with respect to $F_i(x)$, for i=0,1,2,3, we obtain

Table 3.1 Approximate Power of Test using T_4 (σ=1, level of significance 5% m_2=-m_1)

	τ =	.5				1.0				2.0				5.0			
	n =	25	50	100	400	25	50	100	400	25	50	100	400	25	50	100	400
m_1	$\lvert\mu\rvert$																
	.0	1.0	1.0	1.0	1.0	.680	.936	.999	1.0	.493	.643	.826	.998	.971	.999	1.0	1.0
	.2	1.0	1.0	1.0	1.0	.639	.915	.998	1.0	.497	.654	.840	.999	.973	.999	1.0	1.0
	.4	1.0	1.0	1.0	1.0	.508	.818	.995	1.0	.511	.689	.880	1.0	.978	.999	1.0	1.0
1	.6	.996	1.0	1.0	1.0	.300	.555	.960	1.0	.545	.757	.934	1.0	.986	1.0	1.0	1.0
	.8	.713	.980	1.0	1.0	.108	.188	.354	.914	.616	.858	.985	1.0	.993	1.0	1.0	1.0
	1.0	.088	.225	.536	.999	.038	.038	.038	.038	.746	.964	1.0	1.0	.998	1.0	1.0	1.0
	.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	.604	.842	.980	1.0	.743	.893	.983	1.0
	.4	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	.564	.804	.967	1.0	.748	.900	.985	1.0
	.8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	.444	.669	.896	1.0	.767	.920	.991	1.0
2	1.2	1.0	1.0	1.0	1.0	.968	1.0	1.0	1.0	.257	.395	.614	.986	.808	.953	.998	1.0
	1.6	.811	.996	1.0	1.0	.474	.776	.973	1.0	.092	.107	.135	.301	.884	.988	1.0	1.0
	2.0	.002	.002	.002	1.0	.044	.044	.043	1.0	.100	.144	.234	.672	.981	1.0	1.0	1.0

$$\frac{1}{2} = \omega_1\alpha_{10} + \omega_2\alpha_{20} + (1-\omega_1-\omega_2)\alpha_{30}$$

$$\alpha_{01} = \frac{1}{2}\omega_1 + \omega_2\alpha_{21} + (1-\omega_1-\omega_2)\alpha_{31}$$

$$\alpha_{02} = \omega_1\alpha_{12} + \frac{1}{2}\omega_2 + (1-\omega_1-\omega_2)\alpha_{32}$$

$$\alpha_{03} = \omega_1\alpha_{13} + \frac{1}{2}\alpha_{23} + (1-\omega_1-\omega_2)\frac{1}{2}$$

where α_{ab} is defined by (3.1.3). We introduce the symbols

$$A_i = \alpha_{0i} - \alpha_{3i}$$

$$B_i = \alpha_{1i} - \alpha_{3i}$$

$$C_i = \alpha_{2i} - \alpha_{3i}$$

for i=0,1,2,3. After rearrangement, the above equations can be written as:

$$\begin{aligned} A_0 &= \omega_1 B_0 + \omega_2 C_0 \\ A_1 &= \omega_1 B_1 + \omega_2 C_1 \\ A_2 &= \omega_1 B_2 + \omega_2 C_2 \\ A_3 &= \omega_1 B_3 + \omega_2 C_3 \end{aligned} \qquad (3.5.2)$$

A necessary condition for (3.5.2) to have a unique consistent solution for $\{\omega_1,\omega_2\}$ is that any three of the four equations in (3.5.2) have a consistent solution for $\{\omega_1,\omega_2\}$. (A consistent solution is a solution that simultaneously satisfies the designated set of equations.) In other words, each of the following four determinants has value zero,

$$\begin{vmatrix} A_0 & B_0 & C_0 \\ A_1 & B_1 & C_1 \\ A_2 & B_2 & C_2 \end{vmatrix}, \quad \begin{vmatrix} A_0 & B_0 & C_0 \\ A_1 & B_1 & C_1 \\ A_3 & B_3 & C_3 \end{vmatrix}, \quad \begin{vmatrix} A_0 & B_0 & C_0 \\ A_2 & B_2 & C_2 \\ A_3 & B_3 & C_3 \end{vmatrix}, \quad \begin{vmatrix} A_1 & B_1 & C_1 \\ A_2 & B_2 & C_2 \\ A_3 & B_3 & C_3 \end{vmatrix}.$$

Note that, by definition,

$$A_1 - A_0 = B_1 - B_0$$

$$A_2 - A_0 = C_2 - C_0$$

$$B_2 - B_1 = C_2 - C_1$$

from which, our conditions can be simplified as follows:

$$(C_1 - C_0 - A_1 + A_0)[B_0C_2 - B_1(C_0 - A_0) - A_0B_2] = 0$$

$$(A_0 - A_1)[B_0C_2 - B_1(C_0 - A_0) - A_0B_2] = 0$$

$$(A_0 - A_2)[B_0C_2 - B_1(C_0 - A_0) - A_0B_2] = 0$$

$$-(B_2 - B_1)[B_0C_2 - B_1(C_0 - A_0) - A_0B_2] = 0 .$$

Suppose that $B_0C_2 - B_1(C_0 - A_0) - A_0B_2 \neq 0$. Then

$$C_1 - C_0 - A_1 + A_0 = 0$$

$$A_0 - A_1 = 0$$

$$A_0 - A_2 = 0$$

$$B_2 - B_1 = 0 \quad ,$$

it follows that $A_0=A_1=A_2$; $B_0=B_1=B_2$; $C_0=C_1=C_2$. In turn, these give $B_0C_2 - B_1(C_0 - A_0) - A_0B_2 = 0$, which is a contradiction. Hence we must have

$$B_0C_2 - B_1(C_0 - A_0) - A_0B_2 = 0. \qquad (3.5.3)$$

This is therefore a necessary condition for (3.5.2) to have a unique solution for $\{\omega_1,\omega_2\}$.

The reasons that (3.5.2) must have a unique consistent solution for $\{\omega_1, \omega_2\}$ are discussed in the following:

If there exist two sets of solutions for (3.5.2), say $\{\omega_1,\omega_2\}$ and $\{\omega_1',\omega_2'\}$, with $0 \le \omega_1, 0 \le \omega_2, 0 \le \omega_1'$, $0 \le \omega_2'$, $\omega_1+\omega_2<1$, $\omega_1'+\omega_2'< 1$, and either $\omega_1 \neq \omega_1'$, or $\omega_2 \neq \omega_2'$, then we have

$$F_0(x) = \omega_1F_1(x) + \omega_2F_2(x) + (1- \omega_1 - \omega_2)F_3(x)$$

$$= \omega_1'F_1(x) + \omega_2'F_2(x) + (1-\omega_1'- \omega_2')F_3(x) \quad \text{for all } x$$

which implies that

$$0=(\omega_1-\omega_1')[F_1(x)-F_3(x)] + (\omega_2-\omega_2')[F_2(x)-F_3(x)] \text{ , for all } x.$$

From this , it will be shown that the existence of two non-identical solutions $\{\omega_1,\omega_2\}$, $\{\omega_1',\omega_2'\}$ reduces the problem of testing a mixture of three components to one of testing two components. This has already been discussed in sections 3.1 to 3.3, and is not our present interest.

(1) If $\omega_1=\omega_1'$, $\omega_2 \neq \omega_2'$, then $F_2(x)\equiv F_3(x)$, which implies that under H_0, $F_0(x) = \omega_1 F_1(x) + (1- \omega_1)F_2(x)$, a mixture of two components.

(2) If $\omega_1 \neq \omega_1'$, $\omega_2 = \omega_2'$, then similarly $F_0(x) =\omega_2 F_2(x)+(1-\omega_2)F_3(x)$, a mixture of two components.

(3) If $\omega_1 \neq \omega_1'$, $\omega_2 \neq \omega_2'$,

(a) $(\omega_1 -\omega_1')(\omega_2 - \omega_2') > 0$, then

$$F_3(x) = \frac{\omega_1 - \omega_1'}{\omega_1-\omega_1'+\omega_2-\omega_2'} F_1(x) + \frac{\omega_2 - \omega_2'}{\omega_1-\omega_1'+\omega_2-\omega_2'} F_2(x) ,$$

i. e., $F_3(x)$ is a proper mixture of $F_1(x)$ and $F_2(x)$.

(b) $(\omega_1 -\omega_1')(\omega_2- \omega_2') < 0$ and $|\omega_1 - \omega_1'| < |\omega_2- \omega_2'|$

$$F_2(x) = \frac{\omega_1' - \omega_1}{\omega_2 - \omega_2'} F_1(x) + \frac{\omega_1 - \omega_1' +\omega_2 - \omega_2'}{\omega_2 - \omega_2'} F_3(x) ,$$

i. e., $F_2(x)$ is a proper mixture of $F_1(x)$ and $F_3(x)$

(c) $(\omega_1 - \omega_1')(\omega_2 - \omega_2') < 0$ and $|\omega_1 - \omega_1'| > |\omega_2- \omega_2'|$

$$F_1(x) = \frac{\omega_2' - \omega_2}{\omega_1 - \omega_1'} F_2(x) + \frac{\omega_1- \omega_1'+ \omega_2 - \omega_2'}{\omega_1 - \omega_1'} F_3(x) ,$$

i. e., $F_1(x)$ is a proper mixture of $F_2(x)$ and $F_3(x)$.

(d) $\omega_1- \omega_1' = \omega_2' - \omega_2$, then

$$0 = (\omega_1 - \omega_1')F_1(x) + (\omega_2 - \omega_2')F_2(x)= (\omega_1-\omega_1')[F_1(x)-F_2(x)]$$

which implies that $F_1(x) \equiv F_2(x)$.

In each of the above cases (a) to (d), $F_0(x)$ can be reduced to a proper mixture of two components. For example, if $F_3(x)=\omega F_1(x)+(1-\omega)F_2(x)$, then $F_0(x)=\omega_1 F_1(x)+\omega_2 F_2(x)+(1-\omega_1-\omega_2)F_3(x)$

$$=[\omega_1+\omega(1-\omega_1-\omega_2)]F_1(x) + [\omega_2+(1-\omega)(1-\omega_1-\omega_2)]F_2(x)$$

a proper mixture of F_1 and F_2. In conclusion, for F_0 to be a genuine mixture of three components, (3.5.2) must possess a unique consistent solution $\{\omega_1,\omega_2\}$, a necessary condition for which is (3.5.3).

Rewrite (3.5.3) in terms of α's

$$0 = (1/2-\alpha_{01})(1/2-\alpha_{32})+(1/2-\alpha_{31})(1/2-\alpha_{20})+(1/2-\alpha_{30})(1/2-\alpha_{12}) \tag{3.5.4}$$

Our test statistics will be derived from (3.5.4). First let us consider problem A1. Besides the random sample from F_0, we have for i=1,2,3, a random sample $X_{i1},---,X_{in_i}$ from F_i. Define

$$T_5 = (1/2-W_{01})(1/2-W_{32})+(1/2-W_{31})(1/2-W_{20})+(1/2-W_{30})(1/2-W_{12}) \tag{3.5.5}$$

where W_{ij} is defined as in (3.1.2). Under H_0, $ET_5=0$. Choose ε such that $0<\varepsilon\le \dfrac{n_i}{n_o+n_1+n_2+n_3} = p_i \le 1-\varepsilon<0$, i=0,1,2,3, then by Theorem 3.1

$$N^{1/2}((W_{01}-1/2),(W_{32}-1/2),(W_{20}-1/2),(W_{31}-1/2),(W_{12}-1/2),(W_{30}-1/2))$$

has an asymptotic normal distribution with mean

$$N^{1/2}((\alpha_{01}-1/2),(\alpha_{32}-1/2),(\alpha_{20}-1/2),(\alpha_{31}-1/2),(\alpha_{12}-1/2),(\alpha_{30}-1/2)) \tag{3.5.6}$$

and variance-covariance matrix $N((\sigma_{ij}))$, i,j=1,2,---,6, where $N=n_0+n_1+n_2+n_3$ and σ_{ij} are listed below. The function $g(x_1,---,x_6)$ $=x_1x_2+x_3x_4+x_5x_6$ is continuous on R^6, hence by Theorem 3.2, NT_5 has an

asymptotic distribution which is that of $Z_1Z_2+Z_3Z_4+Z_5Z_6$ with $(Z_1,---,Z_6)$ having a 6-variate normal distribution with mean vector (3.5.6) and variance-covariance matrix $((\sigma_{ij}))$ as follows:

$$\sigma_{11} = p_0^{-1}[\int F_1^2 dF_0 - \alpha_{10}^2] + p_1^{-1}[\int F_0^2 dF_1 - \alpha_{01}^2]$$

$$\sigma_{22} = p_3^{-1}[\int F_2^2 dF_3 - \alpha_{23}^2] + p_2^{-1}[\int F_3^2 dF_2 - \alpha_{32}^2]$$

$$\sigma_{33} = p_2^{-1}[\int F_0^2 dF_2 - \alpha_{02}^2] + p_0^{-1}[\int F_2^2 dF_0 - \alpha_{20}^2]$$

$$\sigma_{44} = p_3^{-1}[\int F_1^2 dF_3 - \alpha_{13}^2] + p_1^{-1}[\int F_3^2 dF_1 - \alpha_{31}^2]$$

$$\sigma_{55} = p_1^{-1}[\int F_2^2 dF_1 - \alpha_{21}^2] + p_2^{-1}[\int F_1^2 dF_2 - \alpha_{12}^2]$$

$$\sigma_{66} = p_3^{-1}[\int F_0^2 dF_3 - \alpha_{03}^2] + p_0^{-1}[\int F_3^2 dF_0 - \alpha_{30}^2]$$

$$\sigma_{12} = \sigma_{34} = \sigma_{56} = 0$$

$$\sigma_{13} = -p_0^{-1}[\int F_1 F_2 dF_0 - \alpha_{10}\alpha_{20}]$$

$$\sigma_{14} = p_1^{-1}[\int F_0 F_3 dF_1 - \alpha_{01}\alpha_{31}]$$

$$\sigma_{15} = -p_1^{-1}[\int F_0 F_2 dF_1 - \alpha_{01}\alpha_{21}]$$

$$\sigma_{16} = -p_0^{-1}[\int F_1 F_3 dF_0 - \alpha_{10}\alpha_{30}]$$

$$\sigma_{23} = -p_2^{-1}[\int F_0 F_3 dF_2 - \alpha_{02}\alpha_{32}]$$

$$\sigma_{24} = p_3^{-1}[\int F_1 F_2 dF_3 - \alpha_{13}\alpha_{23}]$$

$$\sigma_{25} = p_2^{-1}[\int F_1 F_3 dF_2 - \alpha_{12}\alpha_{32}]$$

$$\sigma_{26} = p_3^{-1}[\int F_0 F_2 dF_3 - \alpha_{02}\alpha_{03}]$$

$$\sigma_{35} = -p_2^{-1}[\int F_0 F_1 dF_2 - \alpha_{02}\alpha_{12}]$$

$$\sigma_{36} = p_0^{-1}[\int F_2 F_3 dF_0 - \alpha_{20}\alpha_{30}]$$

$$\sigma_{45} = -p_1^{-1}[\int F_2F_3dF_1-\alpha_{21}\alpha_{31}]$$

$$\sigma_{46} = p_3^{-1}[\int F_0F_1dF_3-\alpha_{03}\alpha_{13}] \qquad (3.5.7)$$

where $p_i=n_i/N$.

In the following we will derive the asymptotic variance of T_5. Let μ_r' denote the r-th moment about zero and κ_r denote the r-th cumulant. Then $\mu_4' = \kappa_4+4\kappa_3\kappa_1+3\kappa_2^2+6\kappa_2\kappa_1^2+\kappa_1^4$. Denote $\mu'_{1111}=E(Z_1Z_2Z_3Z_4)$, then μ'_{1111} can be derived formally as follows (David et al. 1966):

Write μ_4' formally as

$$\mu'(r^4) = \kappa(r^4)+4\kappa(r^3)\kappa(r)+3(\kappa(r^2))^2+6\kappa(r^2)(\kappa(r))^2+(\kappa(r))^4$$

and operate with $s\frac{\partial}{\partial r}$ on both sides, after cancelling the factor 4 on both sides, we have

$$\mu'(r^3s) = \kappa(r^3s)+3\kappa(r^2s)\kappa(r)+\kappa(r^3)\kappa(s)+3\kappa(r^2)\kappa(rs)+3\kappa(rs)(\kappa(r))^2$$
$$+3\kappa(r^2)\kappa(r)\kappa(s)+(\kappa(r))^3\kappa(s).$$

Next operate on both sides first with $t\frac{\partial}{\partial r}$, then with $v\frac{\partial}{\partial r}$. After cancelling certain factors on both sides and putting r=s=t=v=1 , we have

$$\mu'_{1111} = \kappa_{1111}+\kappa_{0111}\kappa_{1000}+\kappa_{1110}\kappa_{0001}+\kappa_{1101}\kappa_{0010}+\kappa_{1011}\kappa_{0100}+\kappa_{0011}\kappa_{1100}$$
$$+\kappa_{1010}\kappa_{0101}+\kappa_{1001}\kappa_{0110}+\kappa_{0110}\kappa_{1000}\kappa_{0001}+\kappa_{0101}\kappa_{1000}\kappa_{0010}$$
$$+\kappa_{1100}\kappa_{0010}\kappa_{0001}+\kappa_{0011}\kappa_{1000}\kappa_{0100}+\kappa_{1010}\kappa_{0100}\kappa_{0001}$$
$$+\kappa_{1001}\kappa_{0100}\kappa_{0010}+\kappa_{1000}\kappa_{0100}\kappa_{0010}\kappa_{0001},$$

where κ_{abcd} is the 4-variate cumulant. Since (Z_1,Z_2,Z_3,Z_4) has a 4-variate normal distribution, we have

$$\kappa_{1111}=\kappa_{0111}=\kappa_{1110}=\kappa_{1101}=\kappa_{1011}=0 ,$$

$$\kappa_{1000}=\alpha_{01}-1/2 , \kappa_{0100}=\alpha_{32}-1/2 , \kappa_{0010}=\alpha_{20}-1/2 , \kappa_{0001}=\alpha_{31}-1/2 ,$$

$$\kappa_{1100}=0 , \quad \kappa_{1010}=\sigma_{13} , \quad \kappa_{1001}=\sigma_{14} , \quad \kappa_{0110}=\sigma_{23} ,$$

$$\kappa_{0101}=\sigma_{24} , \quad \kappa_{0011}=0 .$$

Therefore

$$E(Z_1Z_2Z_3Z_4) = \sigma_{13}\sigma_{24}+\sigma_{14}\sigma_{23}+\sigma_{23}(\alpha_{01}-1/2)(\alpha_{31}-1/2)+\sigma_{24}(\alpha_{01}-1/2)(\alpha_{20}-1/2)$$
$$+\sigma_{13}(\alpha_{32}-1/2)(\alpha_{31}-1/2)+\sigma_{14}(\alpha_{32}-1/2)(\alpha_{20}-1/2)$$
$$+(\alpha_{01}-1/2)(\alpha_{20}-1/2)(\alpha_{32}-1/2)(\alpha_{31}-1/2)$$

Since $E(Z_1Z_2)=\sigma_{12}=0$, $E(Z_3Z_4)=\sigma_{34}=0$, we have $Cov(Z_1Z_2,Z_3Z_4)=E(Z_1Z_2Z_3Z_4)$.

Similarly we have

$$Cov(Z_1Z_2,Z_5Z_6)=E(Z_1Z_2Z_5Z_6)$$
$$= \sigma_{15}\sigma_{26}+\sigma_{16}\sigma_{25}+\sigma_{25}(\alpha_{01}-1/2)(\alpha_{30}-1/2)+\sigma_{26}(\alpha_{01}-1/2)(\alpha_{12}-1/2)$$
$$+\sigma_{15}(\alpha_{32}-1/2)(\alpha_{30}-1/2)+\sigma_{16}(\alpha_{32}-1/2)(\alpha_{12}-1/2)$$
$$+(\alpha_{01}-1/2)(\alpha_{12}-1/2)(\alpha_{32}-1/2)(\alpha_{30}-1/2)$$

and

$$Cov(Z_3Z_4,Z_5Z_6)=E(Z_3Z_4Z_5Z_6)$$
$$= \sigma_{35}\sigma_{46}+\sigma_{36}\sigma_{45}+\sigma_{45}(\alpha_{20}-1/2)(\alpha_{30}-1/2)+\sigma_{46}(\alpha_{20}-1/2)(\alpha_{12}-1/2)$$
$$+\sigma_{35}(\alpha_{31}-1/2)(\alpha_{30}-1/2)+\sigma_{36}(\alpha_{31}-1/2)(\alpha_{12}-1/2)$$
$$+ (\alpha_{20}-1/2)(\alpha_{31}-1/2)(\alpha_{12}-1/2)(\alpha_{30}-1/2).$$

By the same method it can be shown that

$$Var(Z_1Z_2) = [\sigma_{11}+(\alpha_{01}-1/2)^2][\sigma_{22}+(\alpha_{32}-1/2)^2],$$

$$Var(Z_3Z_4) = [\sigma_{33}+(\alpha_{20}-1/2)^2][\sigma_{44}+(\alpha_{31}-1/2)^2],$$

$$Var(Z_5Z_6) = [\sigma_{55}+(\alpha_{12}-1/2)^2][\sigma_{66}+(\alpha_{30}-1/2)^2].$$

Finally the asymptotic variance of NT_5 can be derived from the following:

$$Var(Z_1Z_2+Z_3Z_4+Z_5Z_6)$$
$$= Var(Z_1Z_2)+Var(Z_3Z_4)+Var(Z_5Z_6)+2Cov(Z_1Z_2,Z_3Z_4)+2Cov(Z_1Z_2,Z_5Z_6)$$
$$+2Cov(Z_3Z_4,Z_5Z_6). \qquad (3.5.8)$$

From the above argument we see that the asymptotic variance of T_5 (i) is not so easily derived, and (ii) has nonlinear terms. And when we come to derive a consistent estimator for it, we get involve in laborious work and have to work out various estimators for different terms in (3.5.8). The reason for this tedious derivation is that T_5 has terms which are products of random variables. If we look again at the definition of T_5, i.e. (3.5.5), we find that the only possible situation in which T_5 is a linear combination of random variable would be when all the components F_1, F_2, F_3 are known--- viz. problem A3'. In this latter case we can modify T_5 to

$$\tilde{T}_5 = (1/2-R_{01})(1/2-\alpha_{32})+(1/2-\alpha_{31})(1/2-R_{20})+(1/2-R_{30})(1/2-\alpha_{12}) \tag{3.5.9}$$

where R_{0i} is defined as in (3.2.1).

Under H_0, $E\tilde{T}_5=0$,

$$n_0 \mathrm{Var}\ \tilde{T}_5 = -[-(1/2-\alpha_{32})\alpha_{10}+(1/2-\alpha_{31})\alpha_{20}+(1/2-\alpha_{12})\alpha_{30}]^2$$

$$+\int[-(1/2-\alpha_{32})F_1+(1/2-\alpha_{31})F_2+(1/2-\alpha_{12})F_3]dF_0 , \tag{3.5.10}$$

from which a consistent estimator for $n_0 \mathrm{Var}\ \tilde{T}_5$ is therefore

$$n_0 \hat{\mathrm{Var}}\ \tilde{T}_5 = -[-(1/2-\alpha_{32})R_{10}+(1/2-\alpha_{31})R_{20}+(1/2-\alpha_{12})R_{30}]^2$$

$$+ \sum_{i=1}^{n_0} [-(1/2-\alpha_{32})F_1(X_{0i})+(1/2-\alpha_{31})F_2(X_{0i})+(1/2-\alpha_{12})F_3(X_{0i})]^2 \tag{3.5.11}$$

Since $\tilde{T}_5$ has an asymptotic normal distribution, a large sample test can be formulated as follows:

Reject H_0 if $|\tilde{T}_5|(\hat{\mathrm{Var}}\ \tilde{T}_5)^{-1/2} > z_{1-\alpha/2}$.

3.6 Problems A1'. A2', A3'——Four components

Suppose that $X_{i1},\cdots,X_{in_i}$ is a random sample from F_i, i=0,1,2,3,4. Assume that all X's are mutually independent and all F_i are continuous.

We wish to test the following null hypothesis:

$$H_0 : F_0 = \omega_1 F_1 + \omega_2 F_2 + \omega_3 F_3 + (1-\omega_1-\omega_2-\omega_3)F_4 \text{ properly.}$$

Integrating both sides of H_0 with respect to F_i, we have

$$\alpha_{0i} = \omega_1\alpha_{1i} + \omega_2\alpha_{2i} + \omega_3\alpha_{3i} + (1-\omega_1-\omega_2-\omega_3)\alpha_{4i}\ , \quad i=0,1,2,3,4, \qquad (3.6.1)$$

where α_{ab} is defined by (3.1.3). Denote by

$$\begin{aligned} A_i &= \alpha_{0i} - \alpha_{4i} \\ B_i &= \alpha_{1i} - \alpha_{4i} \\ C_i &= \alpha_{2i} - \alpha_{4i} \\ D_i &= \alpha_{3i} - \alpha_{4i} \end{aligned} \qquad (3.6.2)$$

$i=0,1,2,3,4$. (3.6.1) can be rewritten as

$$A_i = \omega_1 B_i + \omega_2 C_i + \omega_3 D_i \qquad i=0,1,2,3,4.$$

Following the same argument as in section 3.5 by setting each of five 4×4 determinants to zero, we have

$$\begin{aligned}
0&=[(A_3-A_0)(C_1-C_2)-(A_2-A_0)(D_1-D_3)-(A_1-A_0)(C_3-C_2)][-B_1D_2+C_2(D_1-D_3)+D_3B_2] \\
0&=[(A_2-A_0)(B_3-B_1)-(B_0-B_1)(D_2-D_3)-(A_3-A_0)(B_2-B_1)][C_2A_3-D_3(A_2-A_0)-A_0C_3] \\
0&=[(C_0-C_2)(D_3-D_1)+(A_1-A_0)(C_3-C_2)-(A_3-A_0)(C_1-C_2)][A_0B_3+D_3(A_1-A_0)-A_3B_1] \\
0&=[(D_1-D_3)(A_2-A_0)-(A_1-A_0)(D_2-D_3)-(D_0-D_3)(B_2-B_1)][-B_1A_2-A_0B_2+C_2(A_1-A_0)] \\
0&=[(D_3-D_1)(A_0-A_2)-(A_3-A_0)(C_1-C_2)-(A_1-A_0)(D_2-D_3)][(B_3-D_3)(C_0-A_0) \\
&\qquad -(C_3-A_3)(B_0-A_0)+(C_1-A_1)(D_0-A_0)]
\end{aligned} \qquad (3.6.3)$$

By definition we have the following relations:

$$\begin{aligned} A_0-A_1 &= B_0-B_1 \\ A_0-A_2 &= C_0-C_2 \\ A_0-A_3 &= D_0-D_3 \\ B_1-B_2 &= C_1-C_2 \end{aligned}$$

$$B_1-B_3 = D_1-D_3$$

$$C_2-C_3 = D_2-D_3 \quad .$$

From these relations it follows that the first factor of each equation in (3.6.3) is equal to each other. Suppose that this common factor is not zero, then the following equations must hold simultaneously:

$$\begin{aligned}
&-B_1D_2+C_2(D_1-D_3)+D_3B_2 = 0 \\
&C_2A_3-D_3(A_2-A_0)-A_0C_3 = 0 \\
&A_0B_3+D_3(A_1-A_0)-B_1A_3 = 0 \qquad (3.6.4) \\
&-B_1A_2-A_0B_2+C_2(A_1-A_0) = 0 \\
&(B_3-D_3)(C_0-A_0)-(C_3-A_3)(B_0-A_0)+(C_1-A_1)(D_0-A_0) = 0 \quad .
\end{aligned}$$

But

$$\begin{aligned}
&(A_3-A_0)(C_1-C_2)-(A_2-A_0)(D_1-D_3)-(A_1-A_0)(C_3-C_2) \\
&= [-B_1D_2+C_2(D_1-D_3)+D_3B_2] - [C_2A_3-D_3(A_2-A_0)-A_0C_3] \\
&\quad + [A_0B_3+D_3(A_1-A_0)-B_1A_3] - [-B_1A_2-A_0B_2+C_2(A_1-A_0)] \\
&\quad + [(B_3-D_3)(C_0-A_0)-(C_3-A_3)(B_0-A_0)+(C_1-A_1)(D_0-A_0)]
\end{aligned}$$

which is equal to zero by (3.6.4), and this is a contradiction of our previous assumption. Therefore when H_0 is true we must have

$$(A_3-A_0)(C_1-C_2)-(A_2-A_0)(D_1-D_3)-(A_1-A_0)(C_3-C_2) = 0 \quad . \qquad (3.6.5)$$

Or in terms of α's

$$\begin{aligned}
&\alpha'_{30}\alpha'_{21}+\alpha'_{31}\alpha'_{02}+\alpha'_{32}\alpha'_{10}-\alpha'_{40}\alpha'_{21}-\alpha'_{41}\alpha'_{02}-\alpha'_{42}\alpha'_{10}-\alpha'_{40}\alpha'_{13}-\alpha'_{41}\alpha'_{30}-\alpha'_{43}\alpha'_{10} \\
&-\alpha'_{40}\alpha'_{32}-\alpha'_{42}\alpha'_{03}+\alpha'_{43}\alpha'_{02}-\alpha'_{41}\alpha'_{23}-\alpha'_{42}\alpha'_{31}+\alpha'_{43}\alpha'_{21} = 0 \qquad (3.6.6)
\end{aligned}$$

where $\alpha'_{ij} = 1/2-\alpha_{ij}$. Thus an appropriate test statistic can be defined as follows:

$$T_6 = W'_{30}W'_{21}+W'_{31}W'_{02}+W'_{32}W'_{10}-W'_{40}W'_{21}-W'_{41}W'_{02}-W'_{42}W'_{10}-W'_{40}W'_{13}-W'_{41}W'_{30}$$
$$-W'_{43}W'_{10}-W'_{40}W'_{32}-W'_{42}W'_{03}+W'_{43}W'_{02}-W'_{41}W'_{23}-W'_{42}W'_{31}+W'_{43}W'_{21} \qquad (3.6.7)$$

where $W'_{ij}=1/2-W_{ij}$, and W_{ij} is defined by (3.1.2). Under H_0, $ET_6=0$ (by independence). The asymptotic distribution of T_6 can be derived similarly to that of T_5 as in section 3.5, but it is difficult to identify unless all the component distributions are known—— viz. we are dealing with problem A3'. In this latter case a modified statistic would be

$$\tilde{T}_6 = R'_{30}\alpha'_{21}+\alpha'_{31}R'_{02}-R'_{40}\alpha'_{21}-\alpha'_{41}R'_{02}-\alpha'_{42}R'_{10}-R'_{40}\alpha'_{13}-\alpha'_{41}R'_{30}-\alpha'_{43}R'_{10}$$
$$+\alpha'_{32}R'_{10}-R'_{40}\alpha'_{32}-\alpha'_{42}R'_{03}+\alpha'_{43}R'_{02}-\alpha'_{41}\alpha'_{23}-\alpha'_{42}\alpha'_{31}+\alpha'_{43}\alpha'_{21} \qquad (3.6.8)$$

where $R'_{ij}=1/2-R_{ij}$ and R_{ij} is defined by (3.2.1).

Under H_0, $E\tilde{T}_6=0$,

$$n_0 \mathrm{Var}\ \tilde{T}_6 = -[a\alpha_{10}+b\alpha_{20}+c\alpha_{30}+d\alpha_{40}]^2 + \int[aF_1+bF_2+cF_3+dF_4]^2 dF_0 \qquad (3.6.9)$$

where

$$a = \alpha'_{32} - \alpha'_{42} - \alpha'_{43}\ ,\quad b = -\alpha'_{31} + \alpha'_{41} - \alpha'_{43}$$

$$c = \alpha'_{21} - \alpha'_{41} + \alpha'_{42}\ ,\quad d = -\alpha'_{21} - \alpha'_{32} - \alpha'_{13}\ .$$

Therefore a consistent estimator for $n_0 \mathrm{Var}\ \tilde{T}_6$ would be

$$n_0 \hat{\mathrm{Var}}\ \tilde{T}_6 = -[aR_{10}+bR_{20}+cR_{30}+dR_{40}]^2 + \sum_{i=1}^{n_0} [aF_1(X_{0i})+bF_2(X_{0i})$$
$$+cF_3(X_{0i})+dF_4(X_{0i})]^2 \qquad (3.6.10)$$

and a large sample test would be

$$\text{Reject } H_0 \text{ if } |T_6|(\hat{\mathrm{Var}}\ \tilde{T}_6)^{-1/2} > z_{1-\alpha/2}\ . \qquad (3.6.11)$$

3.7 Discussion and A Conjecture

In the usual two sample tests, an interesting hypothesis is

$$H_0 : F_0(x) = F_1(x)$$

where F_0 and F_1 are the two underlying distributions to be tested, each of

them is either known or to be estimated from a given sample. Denote by

$$\alpha_{01} = \int F_0 dF_1 .$$

Assume that both F_0, F_1 are continuous. Under H_0 we have

$$\alpha_{01} - 1/2 = 0 \qquad (3.7.1)$$

We may use the statistic

$$W_{01} - 1/2$$

to test H_0, where W_{01} is defined by (3.1.2).

In Table 3.2, the necessary conditions derived from previous sections when testing mixtures of 2, 3, and 4 components are tabulated. For completeness, we also include k=1 by regarding it as a 'one component mixture'. These conditions are readily seen to follow a certain pattern, such that the condition of k components can be derived directly from that of (k-1) components without going through tedious manipulation. The rules for such derivations can be summarized in the follwoing algorithm:

1. For k even (integer), generate new terms through substituting the index i by k every where in the necessary condition of (k-1) component, i=0, 1, ---, k-1, then subtract from the (k-1)-component condition all the new terms generated.

2. For k odd (integer), first group all terms in the condition of (k-1)-components in the manner shown in Table 3.1, (This can also be done by preserving the order of terms when (k-1)-component condition is derived from that of (k-2)-components.) Next, multiphy each group by α_{ik}', where i is the index missed in the group (each group has exactly one such index). Then substract all (new) terms generated from that of (k-1)-components. A (conjectured) 5-component condition is derived in this way. It <u>is</u> only a conjecture and has not been derived analytically.

Table 3.2

Number of components k	Necessary conditions	Total number of terms in the conditions
1	$\alpha'_{01}=0$	$\binom{2}{2}=1$
2	$\alpha'_{01}-\alpha'_{21}-\alpha'_{02}=0$	$\binom{3}{2}=3$
3	$\alpha'_{01}\alpha'_{23}-\alpha'_{21}\alpha'_{03}-\alpha'_{02}\alpha'_{13}=0$	$\frac{1}{2!}\binom{4}{2}\binom{2}{2}=3$
4	$(\alpha'_{01}\alpha'_{23}-\alpha'_{21}\alpha'_{03}-\alpha'_{02}\alpha'_{13})-(\alpha'_{41}\alpha'_{23}-\alpha'_{21}\alpha'_{43}-\alpha'_{42}\alpha'_{13})-(\alpha'_{04}\alpha'_{23}-\alpha'_{24}\alpha'_{03}-\alpha'_{02}\alpha'_{43})$ $-(\alpha'_{01}\alpha'_{43}-\alpha'_{41}\alpha'_{03}-\alpha'_{04}\alpha'_{13})-(\alpha'_{01}\alpha'_{24}-\alpha'_{21}\alpha'_{04}-\alpha'_{02}\alpha'_{14})=0$	$\frac{1}{2!}\binom{5}{2}\binom{3}{2}=15$
5	$(\alpha'_{01}\alpha'_{23}-\alpha'_{21}\alpha'_{03}-\alpha'_{02}\alpha'_{13})\alpha'_{45}-(\alpha'_{41}\alpha'_{23}-\alpha'_{21}\alpha'_{43}-\alpha'_{42}\alpha'_{13})\alpha'_{05}$ $-(\alpha'_{04}\alpha'_{23}-\alpha'_{24}\alpha'_{03}-\alpha'_{02}\alpha'_{43})\alpha'_{15}-(\alpha'_{01}\alpha'_{43}-\alpha'_{41}\alpha'_{03}-\alpha'_{04}\alpha'_{13})\alpha'_{25}$ $-(\alpha'_{01}\alpha'_{24}-\alpha'_{21}\alpha'_{04}-\alpha'_{02}\alpha'_{14})\alpha'_{35}=0$	$\frac{1}{3!}\binom{6}{2}\binom{4}{2}\binom{2}{2}=15$

Note: 1. $\alpha'_{ab} = 1/2 - \alpha_{ab}$.

2. The necessary condition for five component mixture is only a conjecture, it had not been derived analytically.

Note that for the case of k-components, with k=2,3,or 4, the necessary condition is one that includes all the possible combinations of pairs of different indices out of the (k+1) indices, 0, 1,2, ---, k. This can also be seen from the last column of Table 3.2.

3.8 Problem E1

In this section we study the problem of reducing the number of components of a given finite proper mixture by considering two special cases, viz., that of reducing from (1) four components to three, and (2) three components to two. The methods used to derive the test statistics in both cases are the same; they are included here for the purpose of completeness, and for the reason that if these two tests are applied consecutively, we would be able to test for reduction from four components to two components.

In each case, we first derive from the null hypothesis necessary conditions, and then test statistics. As before, we may suppose that each component distribution is either known or there is a sample from it. If not all component distributions are known, the asymptotic distribution of the test statistic is difficult to identify; so from henceforward we will suppose that each component is known.

3.8.1 Reducing Four Components to Three

Let X_1, ---, X_n be a random sample from F_0 and suppose that F_1, F_2, F_3, F_4 are known, continuous. Also assume that F_0 is continuous and that

$$F_0 = \omega_1 F_1 + \omega_2 F_2 + \omega_3 F_3 + (1 - \omega_1 - \omega_2 - \omega_3) F_4 \text{ properly.} \qquad (3.8.1)$$

We wish to test the following null hypothesis:

$$H_0: \quad 1 - \omega_1 - \omega_2 - \omega_3 = 0$$

Define α_{ij} as in (3.1.3) and A_i, B_i, C_i, D_i as in (3.6.2). Under H_0, we have

$$F_0 = \omega_1 F_1 + \omega_2 F_2 + (1 - \omega_1 - \omega_2) F_3 \qquad \text{properly.} \qquad (3.8.2)$$

Integrating both sides of (3.8.2) with respect to F_i, i=0, 1, 2, 3, 4, we obtain

$$A_i = \omega_1 B_i + \omega_2 C_i \ . \qquad (3.8.3)$$

As discussed in section 3.5, when (3.8.2) holds, then necessarily

$$B_0 C_2 - B_1 (C_0 - A_0) - A_0 B_2 = 0 \ . \qquad (3.5.3)$$

If this is so, then the first four equations in (3.8.3) have a unique consistent solution for $\{\omega_1, \omega_2\}$. Let this solution be

$$\hat{\omega}_1 = (A_1 C_2 - A_2 C_1)(B_1 C_2 - B_2 C_1)^{-1}$$

$$\hat{\omega}_2 = (A_1 B_2 - A_2 B_1)(B_2 C_1 - B_1 C_2)^{-1} \ .$$

From the assumption (3.8.1) and H_0, $\hat{\omega}_1$, $\hat{\omega}_2$ must also satisfy the fifth equation in (3.8.3), i. e. we must have

$$A_4 = (A_1 C_2 - A_2 C_1)(B_1 C_2 - B_2 C_1)^{-1} B_4 + (A_1 B_2 - A_2 B_1)(B_2 C_1 - B_1 C_2) C_4 \ ,$$

or

$$0 = \alpha_{04}(B_1 C_2 - B_2 C_1) + \alpha_{01}(B_2 C_4 - C_2 B_4) + \alpha_{02}(C_1 B_4 - B_1 C_4)$$
$$- \alpha_{34}(B_1 C_2 - B_2 C_1) - \alpha_{31}(B_2 C_4 - C_2 B_4) - \alpha_{31}(C_1 B_4 - B_1 C_4) \qquad (3.8.4)$$

The right hand side of this equation is a linear combination of α_{04}, α_{01}, and α_{02} (the only non-constant terms). (3.5.3) can be rewritten as

$$0 = \alpha_{10} C_2 - \alpha_{20} B_1 - \alpha_{30}(C_2 - B_2) + \tfrac{1}{2}(B_1 - B_2) \qquad (3.8.5)$$

In the above argument we find that (3.8.4) and (3.8.5) are two necessary conditions for both assumption (3.8.1) and H_0 to hold. Therefore we may

define test statistics

$$T_7 = R_{10}C_2 - R_{20}B_1 - R_{30}(C_2 - B_2) + \frac{1}{2}(B_1 - B_2) \qquad (3.8.6)$$

$$T_8 = (R_{04} - \alpha_{34})(B_1C_2 - B_2C_1) + (R_{01} - \alpha_{31})(B_2C_4 - C_2B_4)$$
$$+ (R_{02} - \alpha_{32})(C_1B_4 - B_1C_4) , \qquad (3.8.7)$$

where R_{ij} is defined in (3.2.1). Obviously expectations of T_7 and T_8 are equal to the right hand sides of equations (3.8.5) and (3.8.4) respectively, and Var T_7, Var T_8 and Cov(T_7, T_8) can be calculated by using (3.2.3) and (3.3.2) as in the following

$$n \text{ Var } T_7 = -[C_2\alpha_{10} - B_1\alpha_{20} - (C_2 - B_2)\alpha_{30}]^2$$
$$+ \int [C_2F_1 - B_2F_2 - (C_2 - B_2)F_3]^2 dF_0 = t_{11} \quad , \quad \text{say,}$$

$$n \text{ Var } T_8 = -[(B_1C_2 - B_2C_1)\alpha_{04} + (B_2C_4 - C_2B_4)\alpha_{01} + (C_1B_4 - B_1C_4)\alpha_{02}]^2$$
$$+ \int [(B_1C_2 - B_2C_1)F_4 + (B_2C_4 - C_2B_4)F_1 + (C_1B_4 - B_1C_4)F_2]^2 dF_0$$
$$= t_{22} \quad , \quad \text{say,}$$

$$n \text{ Cov}(T_7, T_8) = -[C_2\alpha_{10} - B_1\alpha_{20} - (C_2 - B_2)\alpha_{30}][(B_1C_2 - B_2C_1)\alpha_{04} + (B_2C_4 - C_2B_4)\alpha_{01}$$
$$+ (C_1B_4 - B_1C_4)\alpha_{02}] + \int [C_2F_1 - B_1F_2 - (C_2 - B_2)F_3] \cdot$$
$$[(B_1C_2 - B_2C_1)F_4 + (B_2C_4 - C_2B_4)F_1 + (C_1B_4 - B_1C_4)F_2] dF_0$$
$$= t_{12} \quad , \quad \text{say} \; .$$

t_{11}, t_{22}, and t_{12} can be estimated consistently term by term by Proposition 3.1. We will denote these consistent estimators by $\hat{t}_{11}$, $\hat{t}_{22}$, and $\hat{t}_{12}$ respectively. After rearranging, T_7 and T_8 can each be expressed as means of n i.i.d. random variables, i.e.

$$T_7 = \frac{1}{n} \sum_{k=1}^{n} [C_2F_1(X_k) - B_1F_2(X_k) - (C_2 - B_2)F_3(X_k) + (B_1 - B_2)/2] \qquad (3.8.8)$$

$$T_8 = \frac{1}{n} \sum_{k=1}^{n} [(B_1C_2-B_2C_1)(1-F_4(X_k)-\alpha_{34}) + (B_2C_4-B_4C_2)(1-F_1(X_k)-\alpha_{31}) + (C_1B_4-B_4C_1)(1-F_2(X_k)-\alpha_{32})] \quad (3.8.9)$$

Hence if $nVar\ T_7 < \infty$, by the central limit theorem,

$$T_7(Var\ T_7)^{-1/2} \to N(0,1) \text{ in distribution.}$$

And if $nVar\ T_8 < \infty$,

$$T_8(Var\ T_8)^{-1/2} \to N(0,1) \text{ in distribution.}$$

Let $\underset{\sim}{T} = (T_7,T_8)$, then $\underset{\sim}{T}$ has an asymptotic bivariate normal distribution.

Define

$$Q_n = n\ \underset{\sim}{T} \begin{pmatrix} t_{11} & t_{12} \\ t_{12} & t_{22} \end{pmatrix}^{-1} \underset{\sim}{T}' ,$$

then Q_n has an asymptotic chi-square distribution with 2 degrees of freedom (Wilks, 1962, pp. 261). Applying Theorems 2.1 and 2.2, Q_n and

$$\hat{Q}_n = n\ \underset{\sim}{T} \begin{pmatrix} \hat{t}_{11} & \hat{t}_{12} \\ \hat{t}_{12} & \hat{t}_{22} \end{pmatrix}^{-1} \underset{\sim}{T}' \quad (3.8.10)$$

have the same asymptotic chi-square distribution with 2 degrees of freedom. Under H_0, $ET_7 = ET_8 = 0$, so $\hat{Q}_n$ has an asymptotic central chi-square distribution. Otherwise $\hat{Q}_n$ has an asymptotic noncentral chi-square distribution with 2 degrees of freedom and noncentrality parameter

$$n(\underset{\sim}{ET}) \begin{pmatrix} t_{11} & t_{12} \\ t_{12} & t_{22} \end{pmatrix}^{-1} (\underset{\sim}{ET})'. \quad (3.8.11)$$

Therefore a large sample test is

reject H_0 if $\hat{Q}_n$ is too large. (3.8.12)

3.8.2 Reducing Three Components to Two

Let $X_1,---,X_n$ be a random sample from F_0 and suppose that F_1, F_2, F_3 are known, continuous. Also assume that F_0 is continuous and

$$F_0 = \omega_1 F_1 + \omega_2 F_2 + (1-\omega_1-\omega_2)F_3 \text{ properly.} \tag{3.8.13}$$

We wish to test the following null hypothesis:

$$H_0 : 1-\omega_1-\omega_2 = 0$$

Under H_0,

$$F_0 = \omega_1 F_1 + (1-\omega_1)F_2 \text{ properly.} \tag{3.8.14}$$

Integrating both sides of (3.8.14) with respect to F_i, i=0,1,2,3, we have

$$A_i = \omega_1 B_i . \tag{3.8.15}$$

When H_0 is true, the following condition must hold :

$$\alpha_{10} + \alpha_{02} - \alpha_{12} - 1/2 = 0, \tag{3.1.5}$$

from which it follows that the first three equations in (3.8.15) have a unique consistent solution of ω_1. Let this solution be

$$\hat{\omega}_1 = A_2 / B_2 = (\alpha_{02}-1/2)/(\alpha_{12}-1/2) .$$

Substituting $\hat{\omega}_1$ into the last equation in (3.8.15), we have

$$(\alpha_{03}-\alpha_{23})(\alpha_{12}-1/2) = (\alpha_{13}-\alpha_{23})(\alpha_{02}-1/2) \tag{3.8.16}$$

Define T_4 as in (3.3.1) and

$$T_9 = (R_{03}-\alpha_{23})(\alpha_{12}-1/2) - (\alpha_{13}-\alpha_{23})(R_{02}-1/2) , \tag{3.8.17}$$

where R_{ij} is defined by (3.2.1). (3.1.5) and(3.8.16) are two necessary conditions when the assumption (3.8.13) and H_0 both hold. nVar T_4 was derived in (3.3.3) and will be denoted by s_{11}. While

$$\text{nVar } T_9 = -[(\alpha_{12}-1/2)\alpha_{30}+(\alpha_{13}-\alpha_{23})\alpha_{20}]^2 + [(\alpha_{12}-1/2)F_3+(\alpha_{13}-\alpha_{23})F_2]^2 dF_0$$

$$= s_{22} , \text{ say,}$$

$$\text{nCov}(T_4,T_9) = -(\alpha_{10}-\alpha_{20})[(\alpha_{12}-1/2)\alpha_{30}+(\alpha_{13}-\alpha_{23})\alpha_{20}]$$

$$+\int (F_1-F_2)[(\alpha_{12}-1/2)F_3+(\alpha_{13}-\alpha_{23})F_2]dF_0$$

$$= s_{12}, \text{ say.}$$

Similarly we can estimate s_{11}, s_{12}, s_{22} consistently by Proposition 3.1,

and we will denote these estimators by $\hat{s}_{11}$, $\hat{s}_{12}$, $\hat{s}_{22}$ respectively. Following the same argument as in section 3.8.1, we may use

$$\hat{P}_n = n(T_4,T_9) \begin{pmatrix} \hat{s}_{11} & \hat{s}_{12} \\ \hat{s}_{12} & \hat{s}_{22} \end{pmatrix}^{-1} (T_4,T_9)' \qquad (3.8.18)$$

as a test statistic and

reject H_0 if $\hat{P}_n$ is too large. (3.8.19)

CHAPTER IV

A SIMULATION STUDY

In this chapter we study a new statistic which can be used to test hypotheses of proper mixtures when the components are known, e.g. problems A3, B1 and D2. A computational algorithm is constructed and properties of this kind of statistics are discussed. Then we use simulation procedure for three special cases.

4.1 Outline

Let $X_1,---,X_n$ be a randon sample from F_0 and suppose that F_1 and F_2 are known, continuous and

$$F_1(x) \geq F_2(x) \quad \text{for all } x. \tag{4.1.1}$$

We wish to test the following null hypothesis

$$H_0 : F_0=\omega F_1+(1-\omega)F_2 \text{ properly.} \tag{4.1.2}$$

Define a statistic

$$D = \inf_{0\leq\omega\leq1} \sup_x |F_{0n}(x) - F_0(x)|$$

$$= \inf_{0\leq\omega\leq1} \sup_x |F_{0n}(x) - \omega F_1(x) - (1-\omega)F_2(x)| \tag{4.1.3}$$

where $F_{0n}(x)$ is the empirical distribution function corresponding to $X_1,---,X_n$. D can be rewritten as

$$D = \inf_{0\leq\omega\leq1} \max_{j=1,---,n} K_j(\omega) \tag{4.1.4}$$

where

$$K_j(\omega) = \max \left[\left|\frac{j-1}{n} - \omega F_1(X_{(j)}) - (1-\omega)F_2(X_{(j)})\right|,\right.$$

$$\left|\frac{j}{n} - \omega F_1(X_{(j)}) - (1-\omega)F_2(X_{(j)})\right| \qquad (4.1.5)$$

and $X_{(j)}$ is the j-th order statistic of $X_1,---,X_n$.

On the other hand the distribution of D can be expressed as a weighted mean of conditional distributions (Behboodian, 1972) as follows:

$$\Pr\{ D \le d \mid \omega \}$$

$$= \sum_{i=0}^{n} \binom{n}{i} \omega^i (1-\omega)^{n-i} \Pr\{D \le d \mid \text{ i out of n X's are from } F_1 \text{ and the rest from } F_2\}, \qquad (4.1.6)$$

where ω is the probability that X_1 comes from F_1. We will use (4.1.5) to calculate the conditional distributions and then use (4.1.6) to obtain the (unconditional) distribution of D.

4.2 An Algorithm

In this section we describe an algorithm which enables us to compute from a given random sample the values of D and $\hat{\omega}$, the latter minimizing the expression (4.1.3). From (4.1.5) $K_j(\omega)$, as a function of ω, is defined in the following ways:

(1) When $\frac{2j-1}{2n} \le F_2(X_{(j)})$, we have

$$K_j(\omega) = F_2(X_{(j)}) - \frac{j-1}{n} + \omega[F_1(X_{(j)}) - F_2(X_{(j)})], \qquad (4.2.1)$$

an increasing function of ω.

(2) When $F_2(X_{(j)}) < \frac{2j-1}{2n} < F_1(X_{(j)})$, we have

$$K_j(\omega) = \begin{cases} -F_2(X_{(j)}) + \frac{j}{n} + \omega[-F_1(X_{(j)}) + F_2(X_{(j)})] & \text{if } 0 \le \omega < \omega_j^* \\ F_2(X_{(j)}) - \frac{j-1}{n} + \omega[F_1(X_{(j)}) - F_2(X_{(j)})], & \text{if } \omega_j^* \le \omega \le 1 \end{cases} \qquad (4.2.2)$$

with $\omega_j^* = [\frac{2j-1}{2n} - F_2(X_{(j)})][F_1(X_{(j)}) - F_2(X_{(j)})]^{-1}$, decreasing first in the interval $[0,\omega_j^*)$, then increasing in the interval $[\omega_j^*,1]$. (4.2.3)

(3) When $F_1(X_{(j)}) \le \frac{2j-1}{2n}$, we have

$$K_j(\omega) = \frac{j}{n} - F_2(X_{(j)}) + \omega[-F_1(X_{(j)})+F_2(X_{(j)})], \quad (4.2.4)$$

a decreasing function of ω.

It can be seen from the above argument that, for each j=1, ---, n, $K_j(\omega)$ is either a straight line segment or a sequence of connected straight lines. And from (4.1.4) we see that the value of D is attained when $\hat{\omega}$ is the minimax point among all the intersection points of straight lines $K_j(\omega)$, j=1,---,n. Using this fact, we construct a computational algorithm for finding the values of D and $\hat{\omega}$ when a random sample is given. (Programs written in Fortran IV are available from the author.)

Step 1 : For each j=1,---,n , let a(j)=1,2 or 3 according to whether $K_j(\omega)$ is of case 1,2 or 3 respectively. Let

$$b(0) = \max_{j=1,---,n} K_j(0). \quad (4.2.5)$$

Denote by

$$K_j'(0) = \left.\frac{dK_j(\omega)}{d\omega}\right|_{\omega=0} .$$

Then find j_0 such that

$$K_{j_0}'(0) = \max \{ K_j'(0) : K_j(0)=b(0) \}.$$

If more than one such j_0 exist, choose any one of them,

Step 2 : If $a(j_0)=1$, put

$$D=b(0), \text{ and } \hat{\omega}=0 \quad (4.2.6)$$

and the algorithm stops. Otherwise let $\omega_0=0$ and proceed to step 3.

Step 3 : Compute the intersection points of the line $K_{j_0}(\omega)$ with all the other lines $K_j(\omega)$ and denote these points by (ω_j^1, B_j^1) for $j \neq j_0$. Let

$$\omega_1 = \min_{j \neq j_0} \max (\omega_j^1, \omega_0) \quad (4.2.7)$$

Step 4 : If $\omega_1 \geq 1$, put

$$D=K_{j_0}(1), \text{ and } \hat{\omega}=1 \tag{4.2.8}$$

and stop the algorithm. Otherwise let j_1 be such that

$$K'_{j_1}(\omega_1) = \max \{K'_j(\omega_1): j \neq j_0, \omega_j^1=\omega_1\} . \tag{4.2.9}$$

If more than one such j_1 exist, choose any one of them.

Step 5 : If $a(j_1)=1$, or both $a(j_1)=2$ and $\omega^*_{j_1} < \omega_1$, with ω^*_j defined by (4.2.3), put

$$D=K_{j_1}(\omega_1), \text{ and } \hat{\omega}=\omega_1 , \tag{4.2.10}$$

and stop the algorithm. Otherwise proceed to step 3 by changing indices appropriately.

4.3 Numerical Results

For exploratory purpose three numerical examples are studied by the simulation procedure. Namely, when the component distributions are respectively:

(1) $F_1 \sim N(-1.5,1)$ and $F_2 \sim N(1.5,1)$

(2) $F_1 \sim N(-2,1)$ and $F_2 \sim N(2,1)$

(3) $F_1 \sim E(1.5)$ and $F_2 \sim N(1,1/16)$.

The simulation procedure can be described as follows:

(a). For each i=0,1,---,n, where n is the sample size, a set of random numbers $\{X_1,---,X_i\}$ is generated according to the distribution F_1, and a second set of random numbers $\{X_{i+1},---X_n\}$ is generated according to the distribution F_2. They are then combined to form a sample for F_0, denoted by $s(i|n)$, subject to the condition that i out of the n random variables are distributed as F_1 and the rest are distributed as F_2. Applying the algorithm in

section 4.2 to $s(i|n)$, we then obtain a pair of values of $(\hat{\omega},D)$.

(b) Repeat (a) until the desired number of samples is achieved (200 in our cases) and the pairs of values of $(\hat{\omega},D)$ are computed. From these values of D, we compute the conditional empirical distributions of D, denoted by $F_n(x|i)$, subject to the condition that i out of n X's are distributed as F_1 and the rest are distributed as F_2, i=1,---,n.

(c) From (4.1.6) we calculate for fixed ω the unconditional empirical distribution of D.

For case (1), samples of sizes 10, 20, 40 respectively were generated. The resulting empirical distributions are summarized in Table 4.1.

For cases (2) and (3), samples of sizes 10, 20 respectively were generated. The resulting empirical distributions are summarized in Tables 4.2 and 4.3.

Table 4.1 Empirical distribution function of the statistic D calculated by (4.1.7) when the components are N(-1,5), N(1,5) respectively

Sample size n = 10

ω / d	.0	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
.095	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
.120	.01	.025	.037	.046	.053	.058	.061	.063	.064	.065	.065
.130	.04	.060	.072	.080	.087	.092	.096	.099	.101	.103	.103
.140	.075	.097	.112	.124	.134	.143	.152	.160	.166	.170	.171
.150	.150	.164	.180	.195	.211	.226	.240	.252	.262	.267	.269
.160	.210	.224	.246	.270	.294	.317	.336	.353	.365	.372	.374
.170	.250	.277	.311	.343	.373	.399	.422	.440	.453	.461	.464
.180	.290	.336	.383	.427	.467	.502	.531	.554	.571	.581	.584
.190	.320	.390	.454	.511	.560	.600	.633	.657	.674	.684	.687
.200	.355	.439	.514	.579	.632	.675	.707	.730	.746	.755	.758
.210	.395	.494	.578	.647	.700	.740	.769	.789	.802	.809	.812
.220	.470	.559	.637	.702	.754	.793	.820	.839	.850	.856	.858
.230	.535	.617	.689	.749	.800	.835	.863	.882	.894	.900	.903
.240	.595	.669	.732	.785	.827	.860	.884	.901	.912	.919	.921
.255	.665	.731	.784	.829	.865	.893	.915	.930	.940	.946	.948
.265	.705	.575	.802	.842	.876	.904	.925	.941	.951	.957	.959
.280	.740	.797	.842	.878	.908	.931	.948	.960	.968	.973	.974
.285	.775	.802	.858	.890	.917	.938	.954	.966	.974	.978	.980
.295	.815	.851	.881	.907	.929	.948	.963	.974	.981	.986	.987
.315	.840	.877	.906	.929	.947	.962	.973	.981	.986	.989	.990
.330	.855	.894	.923	.945	.962	.974	.982	.988	.991	.993	.994
.355	.895	.925	.947	.963	.975	.983	.989	.993	.996	.997	.998
.385	.935	.952	.966	.976	.983	.989	.992	.995	.997	.998	.999
.395	.95	.963	.973	.980	.986	.991	.994	.996	.998	.998	.999
.415	.975	.982	.986	.990	.993	.995	.997	.998	.999	.999	.999
.595	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 4.1 (Cont.)

Sample size n = 20

d \ ω	.0	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
.059	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
.094	.015	.047	.067	.082	.092	.098	.102	.105	.108	.110	.111
.108	.10	.134	.179	.214	.238	.251	.261	.269	.276	.282	.284
.122	.220	.288	.355	.402	.433	.453	.468	.481	.492	.500	.503
.129	.275	.361	.437	.488	.522	.546	.566	.582	.594	.602	.605
.143	.425	.504	.587	.650	.694	.724	.743	.755	.763	.767	.769
.164	.580	.661	.737	.792	.831	.859	.878	.890	.897	.901	.903
.185	.725	.779	.838	.880	.907	.924	.936	.944	.951	.954	.956
.199	.785	.833	.879	.914	.937	.953	.963	.970	.975	.978	.979
.227	.880	.913	.938	.957	.971	.981	.988	.993	.995	.997	.997
.262	.955	.966	.977	.986	.992	.995	.998	.999	.999	1.0	1.0
.339	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Sample size n = 40

d \ ω	.0	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
.040	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
.060	.010	.016	.023	.025	.029	.034	.036	.035	.033	.034	.034
.080	.165	.191	.236	.264	.284	.305	.327	.344	.352	.353	.352
.100	.420	.518	.564	.597	.631	.662	.682	.692	.698	.704	.707
.120	.635	.748	.796	.828	.857	.882	.899	.910	.918	.923	.925
.130	.735	.827	.876	.902	.918	.931	.940	.948	.954	.959	.961
.140	.805	.870	.913	.940	.954	.963	.969	.975	.980	.983	.984
.150	.860	.903	.942	.965	.977	.983	.987	.989	.990	.992	.992
.170	.945	.954	.973	.987	.933	.995	.997	.998	.999	1.0	1.0
.190	.965	.972	.983	.992	.996	.998	.999	.999	.999	1.0	1.0
.210	.980	.991	.995	.997	.999	.999	1.0	1.0	1.0	1.0	1.0
.270	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 4.2 Empirical distribution function of the statistic D calculated by (4.1.7) when the components are N(-2,1), N(2,1) respectively

Sample size n = 10

ω / d	.0	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
.085	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
.120	.020	.027	.040	.053	.061	.066	.069	.069	.069	.069	.068
.140	.055	.095	.130	.157	.175	.186	.192	.195	.197	.198	.198
.160	.150	.213	.263	.301	.326	.344	.357	.365	.371	.374	.375
.180	.270	.347	.409	.455	.490	.515	.533	.546	.555	.560	.562
.200	.415	.499	.566	.618	.656	.683	.702	.716	.724	.729	.730
.220	.515	.604	.671	.720	.758	.787	.809	.825	.837	.844	.846
.240	.605	.682	.745	.794	.833	.863	.886	.903	.914	.921	.924
.260	.705	.770	.822	.862	.894	.917	.935	.948	.957	.962	.964
.280	.785	.831	.869	.900	.925	.944	.958	.968	.975	.979	.980
.300	.840	.875	.905	.929	.948	.962	.972	.980	.985	.987	.988
.380	.955	.968	.977	.983	.988	.992	.995	.996	.998	.999	.999
.490	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Sample size n = 20

ω / d	.0	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
.058	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
.094	.050	.076	.082	.084	.087	.090	.094	.098	.102	.105	.106
.112	.150	.228	.258	.268	.271	.275	.282	.292	.303	.311	.314
.130	.280	.397	.456	.487	.504	.515	.528	.543	.558	.569	.573
.148	.445	.564	.629	.669	.696	.715	.713	.746	.758	.766	.769
.166	.620	.719	.771	.806	.828	.844	.856	.867	.877	.884	.886
.184	.740	.819	.860	.886	.905	.919	.930	.938	.943	.947	.948
.202	.815	.877	.909	.930	.947	.959	.968	.974	.978	.980	.981
.220	.875	.925	.949	.964	.975	.984	.989	.993	.994	.995	.995
.238	.905	.963	.975	.983	.990	.994	.997	.999	.999	1.0	1.0
.256	.930	.978	.986	.992	.994	.997	.998	.999	.999	1.0	1.0
.418	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 4.3 Empirical distribution function of the statistic D calculated by (4.1.7) when the components are E(1.5), N(1.1/16) respectively

Sample size n = 10

ω / d	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
.085	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
.120	.01	.018	.029	.041	.052	.059	.058	.052	.042	.034	.035
.160	.175	.215	.260	.299	.322	.327	.318	.299	.271	.235	.205
.200	.420	.501	.565	.619	.657	.675	.670	.638	.576	.488	.405
.240	.600	.712	.771	.813	.841	.853	.847	.821	.773	.699	.605
.280	.800	.855	.885	.905	.919	.925	.922	.904	.868	.815	.755
.320	.870	.915	.937	.948	.955	.958	.954	.942	.918	.885	.840
.360	.920	.955	.967	.972	.973	.973	.972	.967	.956	.938	.910
.400	.970	.985	.988	.987	.989	.990	.991	.991	.988	.978	.960
.440	.985	.992	.995	.996	.996	.996	.995	.994	.992	.986	.985
.560	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Sample size n = 20

ω / d	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
.058	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
.103	.075	.122	.149	.168	.178	.178	.172	.157	.142	.136	.085
.121	.180	.283	.342	.362	.372	.384	.393	.382	.355	.324	.210
.139	.325	.453	.517	.545	.571	.593	.598	.580	.555	.530	.430
.157	.500	.615	.674	.706	.737	.758	.759	.743	.718	.672	.545
.175	.640	.727	.788	.820	.841	.850	.845	.833	.823	.787	.645
.202	.810	.854	.883	.905	.922	.936	.940	.931	.911	.873	.760
.238	.910	.946	.958	.966	.974	.978	.977	.973	.968	.957	.890
.265	.965	.973	.985	.990	.991	.990	.986	.982	.979	.965	.935
.301	.975	.992	.996	.997	.998	.998	.995	.992	.992	.986	.970
.418	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Tables 4.1 – 4.3 show that:

(1) In general, as the sample size n increases, the entire range of the empirical distribution function moves toward zero. This is as expected, since when the hypothesis of mixture is true, the value of D would be close to zero.

(2) For the two normal cases, value in Table 4.1 is less than the corresponding value in Table 4.2 for fixed n, d and ω. This indicates that the statistic D performs better when the two components are further apart.

(3) In all three cases,as the actual mixing proportion ω increases from zero to one, the mean of e.d.f.'s first decreases, then increases The same phenomenon holds for almost all the percentiles of the e.d.f. In other words, the statistic D when used to test the hypothesis of mixture, will perform better when the actual mixing proportion is near .5.

(4) In all three cases, when $.1 < \omega < .9$, the e.d.f.'s reamin relatively constant for varying ω. This is especially true when sample size n becomes larger.

The average values of $\hat{\omega}$ over each 200 samples are summarized in Table 4.4. It appears from Table 4.4 that unless the actual proportion $\omega = i/n$ is close to either 0 or 1, the averaged values of $\hat{\omega}$ are very close to ω. This is an indication that the above procedures can also be used to estimate the true proportion parameter of the mixture.

Using nominal significance levels $\alpha = .025, .05$, we also calculate, from the above values of D, approximate values $\tilde{\alpha}$ of the actual significance levels in Tables 4.5 - 4.7. More explicitly, the $\tilde{\alpha}$'s are calculated by the following formula:

Table 4.4 Average values of $\hat{\omega}$ over 200 generated samples (In cases (1) and (2), when $\omega > .5$, use 1-the value corresponding to $1-\omega$.)

The actual proportion from F_1, ω	Case (1)			Case (2)		Case (3)	
	n = 10	20	40	10	20	10	20
.0	.0555	.0560	.0382	.0635	.0411	.1051	.0880
.025			.0511				
.050		.0701	.0612		.0732		.0987
.075			.0743				
.100	.1072	.1092	.0995	.1182	.1052	.1738	.1445
.125			.1286				
.150		.1575	.1415		.1430		.1763
.175			.1710				
.200	.1964	.1945	.1964	.1926	.1969	.2210	.2012
.225			.2173				
.250		.2384	.2378		.2406		.2576
.275			.2622				
.300	.2908	.2941	.3064	.2948	.3059	.2944	.2967
.325			.3277				
.350		.3498	.3454		.3552		.3443
.375			.3692				
.400	.3909	.4011	.3975	.4085	.4080	.3809	.3739
.425			.4280				
.450		.4510	.4461		.4487		.4352
.500	.5022	.4890	.4977	.4906	.5006	.4739	.4525
.550							.5546
.600						.5930	.6116
.650							.6403
.700						.6703	.6472
.750							.7463
.800						.7308	.7787
.850							.7909
.900						.7768	.8429
.950							.8577
1.000						.8294	.8823

$$\tilde{\alpha} = \sum_{i=0}^{n} \binom{n}{i} \omega^i (1-\omega)^{n-i} \frac{1}{200} (\text{no. of points } (\hat{\omega}, D) \text{ above the } D_{1-\alpha} \text{ curve}$$

$$| \; i \text{ out of } n \text{ X's are distributed as } F_1) \,, \qquad (4.3.1)$$

where the $D_{1-\alpha}$ curve is formed by connecting all those points $(\frac{i}{n}, D_{1-\alpha,i})$ $i=1, ---, n$, with $D_{1-\alpha,i}$ satisfying

$$1 - \alpha \simeq \Pr\{ D \leq D_{1-\alpha,i} \mid \frac{i}{n} \} \qquad (4.3.2)$$

Tables 4.5-4.7 show that :

(1) As sample sizes n increase, in general, the approximate values of the actual significance appear to approach the nominal significance level (α). This fact is as expected, since as n increases, the performance of the test statistic usually improves.

(2) When ω is not too close to either 0 or 1, the values of $\tilde{\alpha}$ remain relatively constant. This is due to the fact that when ω is close to either 0 or 1, the small number of random variates generated for one component causes a great deal of variability among the 200 combined samples, which in turn causes the values of D to become larger. Hence at both ends of [0,1] we tend to have larger values of $\tilde{\alpha}$.

(3) In the cases of two normal components, corresponding values in Tables 4.5 and 4.6 are fairly close to each other. This indicates that if we use the statistic D to test (4.1.2) and reject if D is too large, then this test would be consistent with respect to the distance between the components.

(4) The values of $\tilde{\alpha}$ are closer to the nominal significance levels in both normal cases than in the exponential-normal case.

Table 4.5 Approximate values $\tilde{\alpha}$ of the actual significance levels of the statistic D when the components are N(-1.5,1), N(1.5,1) respectively (When $\omega > .5$, use the values corresponding to $1-\omega$.)

	$\alpha = .05$			$\alpha = .025$		
	n = 10	20	40	n = 10	20	40
ω						
.0	.0700	.0800	.0775	.0325	.0425	.0550
.05	.0591	.0685	.0695	.0287	.0309	.0389
.10	.0501	.0554	.0526	.0236	.0262	.0298
.15	.0420	.0462	.0446	.0205	.0226	.0238
.20	.0349	.0406	.0433	.0166	.0192	.0227
.25	.0291	.0375	.0448	.0132	.0166	.0237
.30	.0248	.0360	.0475	.0104	.0151	.0246
.35	.0220	.0353	.0501	.0084	.0142	.0247
.40	.0203	.0347	.0512	.0070	.0126	.0233
.45	.0195	.0342	.0510	.0062	.0131	.0213
.50	.0192	.0340	.0507	.0060	.0130	.0202

Table 4.6 Approximate values $\tilde{\alpha}$ of the actual significance levels of the statistic D when the components are N(-2,1), N(2,1) respectively (When $\omega > .5$, use the values corresponding to $1-\omega$.)

	$\alpha = .05$		$\alpha = .025$	
	n = 10	20	n = 10	20
ω				
.0	.0825	.0500	.0550	.0350
.05	.0711	.0340	.0420	.0170
.10	.0599	.0291	.0321	.0141
.15	.0497	.0287	.0247	.0144
.20	.0409	.0296	.0192	.0152
.25	.0338	.0309	.0153	.0162
.30	.0285	.0322	.0127	.0173
.35	.0247	.0331	.0112	.0182
.40	.0222	.0335	.0105	.0186
.45	.0209	.0336	.0101	.0188
.50	.0204	.0336	.0101	.0188

Table 4.7 Approximate values $\tilde{\alpha}$ of the actual significance levels of the statistic D when the components are E(1.5), N(1,1/16) respectively

	α = .05		α = .025	
	n = 10	20	n = 10	20
ω				
0	.0925		.0350	
.05	.0667	.0532	.0255	.0267
.1	.0508	.0508	.0202	.0204
.15	.0403	.0467	.0171	.0160
.2	.0331	.0416	.0153	.0131
.25	.0279	.0364	.0139	.0116
.3	.0242	.0318	.0129	.0108
.35	.0215	.0277	.0120	.0101
.4	.0197	.0243	.0111	.0095
.45	.0188	.0222	.0103	.0093
.5	.0186	.0217	.0094	.0099
.55	.0192	.0229	.0090	.0113
.6	.0206	.0253	.0088	.0134
.65	.0227	.0282	.0089	.0154
.7	.0258	.0308	.0096	.0168
.75	.0297	.0327	.0111	.0175
.8	.0345	.0345	.0136	.0183
.85	.0401	.0384	.0176	.0211
.9	.0463	.0476	.0235	.0277
.95	.0529	.0656	.0316	.0376
1.0	.0600	.095	.0425	.0500

CHAPTER V

MISCELLANEOUS PROBLEMS

In this chapter we study various problems which have not been discussed in the previous chapters.

5.1 Problem B2 — Normal Components

Let $X_1,---,X_n$ be a random sample from F_0 and suppose that F_1 and F_2 are of known form, symmetric, and $F_2(x)=F_1(x-t)$, $t>0$ known. Assume also that

$$F_0 = \omega F_1 + (1-\omega)F_2 \quad \text{properly} . \tag{5.1.1}$$

We wish to test the following null hypothesis:

$$H_0 : \omega = 0 \text{ or } 1. \tag{5.1.1}$$

Assume further that F_i is distributed as $N(m_i,\sigma^2)$, $i=1,2$. (note that $m_2-m_1=t>0$.) Behboodian (1972) shows that under (5.1.1), the sample mean $\bar{X}_n = \frac{1}{n}\Sigma X_i$ has density

$$f(x,\omega) = \left(\frac{n}{2\pi}\right)^{1/2}\frac{1}{\sigma}\sum_{i=0}^{n}\binom{n}{i}\omega^i(1-\omega)^{n-i}\exp\{-\frac{n}{2\sigma^2}[x-m_2+\frac{i}{n}t]^2\}. \tag{5.1.3}$$

It follows that, for $0<\omega<1$

$\frac{f(x,\omega)}{f(x,0)}$ is a decreasing function of x, and $\frac{f(x,\omega)}{f(x,1)}$ is an increasing function of x. (5.1.4)

Therefore a test can be formulated as follows:

$$\text{Reject } H_0 \text{ if } b < \bar{X}_n < a \tag{5.1.5}$$

where a, b are chosen such that

$$\int_{-\infty}^{a} f(x,0)dx = \Phi\left(\frac{a-m_2}{\sigma/\sqrt{n}}\right) = 1-\alpha/2 \ , \tag{5.1.6}$$

$$\int_{-\infty}^{b} f(x,1)dx = \Phi\left(\frac{b-m_1}{\sigma/\sqrt{n}}\right) = \alpha/2 \ . \tag{5.1.7}$$

and α is the level of significance.

The above method can be applied to the case when there exists a statistic with density satisfying (5.1.4).

5.2 Problem B3

Let $X_1,---,X_n$ be a random sample from F_0 and suppose that F_1 and F_2 are of known forms, and have densities f_1 and f_2 respectively. Also assume that

$$F_0 = \omega F_1 + (1-\omega)F_2 \quad \text{properly} \ , \tag{5.2.1}$$

then F_0 has density $f_0=\omega f_1+(1-\omega)f_2$. We wish to test the following null hypothesis:

$$H_0 \ : \ \omega \leq \omega_0$$

against the alternative hypothesis

$$H_1 \ : \ \omega \geq \omega_1 \ ,$$

where $0\leq\omega_0<\omega_1\leq 1$. First let us state a lemma.

Lemma 1

Let $X_1,---,X_n$ be a random sample from F_0. Assume that F_0 has density $f_0(x,\omega)$, where ω is a real-valued parameter, and that, for $\omega<\omega'$,

$$\frac{f_0(x,\omega')}{f_0(x,\omega)} \quad \text{is a nondecreasing function of } x. \tag{5.2.2}$$

Then the test

$$\psi(X_1,---,X_n) = \begin{cases} 1 & \text{if } L_n(X_1,---,X_n;\omega_0,\omega_1) > c \\ r & \text{if } L_n(X_1,---,X_n;\omega_0,\omega_1) = c \\ 0 & \text{if } L_n(X_1,---,X_n;\omega_0,\omega_1) < c \end{cases} \tag{5.2.3}$$

where

$$L_n(X_1,---,X_n;\omega,\omega') = \prod_{i=1}^{n} \frac{f_0(X_i,\omega')}{f_0(X_i,\omega)} \quad , \tag{5.2.4}$$

and c and r (0<r<1) are constants determined by $E_{\omega_0}\psi(X_1,---,X_n)=\alpha$, maximizes the minimum power for testing H_0 against H_1.

Proof: See Lehmann (1959), pp. 330.

Lemma 1 can be extended by slightly altering the assumption (5.2.2) as follows:

Lemma 2

If in Lemma 1, instead of (5.5.2), assume that for $\omega < \omega'$ $\dfrac{f_0(x,\omega')}{f_0(x,\omega)}$ is a nondecreasing function of some suitably chosen function $t(x)$. (5.2.5)

Then the test

$$\psi(t(X_1),---,t(X_n)) = \begin{cases} 1 & \text{if } L_n(t(X_1),--,t(X_n);\omega_0,\omega_1) > c \\ r & \text{if } L_n(t(X_1),---,t(X_n);\omega_0,\omega_1) = c \\ 0 & \text{if } L_n(t(X_1),---,t(X_n);\omega_0,\omega_1) < c \end{cases} \tag{5.2.6}$$

where

$$L_n(t(X_1),---,t(X_n);\omega,\omega') = \prod_{i=1}^{n} \frac{f_0(X_i,\omega')}{f_0(X_i,\omega)}$$

and c and r(0<r<1) are constants determined by $E_{\omega_0}\psi(t(X_1),---,t(X_n))=\alpha$, maximizes the minimum power for testing H_0 against H_1.

Proof: A slight modification of Lehmann's proof for Lemma 1.

In the case of a 2-component mixture, the density of F_0 is of the form $f_0=\omega f_1+(1-\omega)f_2$. In order to apply Lemmas 1 or 2, $f_0(x,\omega)$ must satisfy the assumptions (5.2.2) or (5.2.5). Note that we may write

$$\frac{f_0(x,\omega')}{f_0(x,\omega)} = 1+\frac{\omega'-\omega}{\dfrac{f_2(x)}{f_1(x)-f_2(x)}+\omega}$$

$$= 1+\frac{\omega'-\omega}{\left(\dfrac{f_1(x)}{f_2(x)}-1\right)^{-1}+\omega}. \qquad (5.2.7)$$

For $\omega<\omega'$, this is a nondecreasing function of $f_1(x)/f_2(x)$. Hence if we let

$$t(x) = f_1(x)/f_2(x)\ , \qquad (5.2.8)$$

we may apply Lemma 2 to obtain a test (i.e. (5.2.6)) which maximizes the minimum power for testing H_0 against H_1. If further $f_1(x)/f_2(x)$ is a nondecreasing function of another suitably chosen function $t_1(x)$, then $f_0(x,\omega')/f_0(x,\omega)$ is also a nondecreasing function of $t_1(x)$, Lemma 2 can be applied.

5.3 Problems B4 and B5

Let $X_1,---,X_n$ be a random sample from F_0 and suppose that F_1 is known, $F_2(x)=F_1(x-t)$ (in problem B4) or $F_2(x)=F_1(x/t)$ (in problem B5), with t unknown. We wish to test the following null hypothesis:

H_0 : $F_o = \omega F_1 + (1-\omega)F_2$ properly.

Owing to the presence of the nuisance parameter t, we can not apply the methods discussed in sections 2.1, 3.1, 3.2 or 3.3. Behboodian (1976) uses a method of moments to estimate the (unknown) parameters ω and t. We could of course use this estimated value of t, and apply the methods discussed

in the previous sections as if both components were known to have these estimated values to derive test statistics. In the following we will explore another approach to problem B4.

Denote by $F_{0n}(x)$ the empirical distribution function corresponding to the random sample $X_1,---,X_n$. Without loss of generality, in the following we will assume $t>0$, (if $t<0$ the following still holds by exchanging $F_1(x)$ and $F_2(x)$). Then $F_2(x) \le F_1(x)$, and

$$F_0(x) = F_1(x) + (1-\omega)[F_2(x)-F_1(x)] \ge F_1(x). \tag{5.3.1}$$

It follows that for $i=0,1,---,n$,

$$\sup_{X_{(i)} \le x \le X_{(i+1)}} |F_{0n}(x)-F_0(x)| \ge \max \left(\frac{i}{n} - F_1(X_{(i)}),\ F_1(X_{(i+1)}) - \frac{i}{n}\right) ,$$

where $X_{(0)}=-\infty$, $X_{(n+1)}=\infty$. Therefore

$$\sup_{-\infty<x<\infty} |F_{0n}(x)-F_0(x)| \ge \max_{i=1,---,n} \left[\max\left(\frac{1}{n} - F_1(X_{(i)}),\ F_1(X_{(i+1)}) - \frac{1}{n}\right)\right]$$

and then

$$\begin{aligned} D = \inf_{0\le\omega\le1} \sup_{-\infty<x<\infty} |F_{0n}(x)-F_0(x)| &\ge \max_{i=1,---,n} \left[\max\left(\frac{i}{n} - F_1(X_{(i)}),\right.\right. \\ &\qquad \left.\left. F_1(X_{(i+1)}) - \frac{i}{n}\right)\right] \\ &\ge D_n \end{aligned} \tag{5.3.2}$$

where D_n is the Kolmogorov-Smirnov statistic, the distribution of which can be used along with (5.3.2) to obtain a test for H_0.

5.4 Problems C1, C2, C3

Let $X_{01},---,X_{0n_0}$ be a random sample from F_0 and suppose that for $i=1,2$, F_i is continuous and is either known or there is a random sample for it. We wish to test the following null hypothesis:

$$H_0 : F_0 = \omega F_1 + (1-\omega)F_2 \text{ for some } \omega \text{ in } [a,b] \text{ with } 0<a<b<1, \text{ known} . \tag{5.4.1}$$

We show in the following that we can use the test statistics, T_2, T_3, T_4 discussed in sections 3.1 - 3.3 for problems C_1, C_2, C_3 respectively.

Define

$$F_1^* = bF_1 + (1- b)F_2 \ ,$$

$$F_2^* = aF_1 + (1- a)F_2 \qquad (5.4.2)$$

Substituting these expressions into (5.4.1) and simplifying we may express F_0 as

$$F_0 = \frac{\omega - a}{b - a} F_1^* + \frac{b - \omega}{b - a} F_2^*$$

Let

$$\omega^* = \frac{\omega- a}{b- a} \ , \qquad (5.4.3)$$

then $F_0 = \omega^* F_1^* + (1 - \omega^*)F_2^*$, with $0 \le \omega^* \le 1$. Since both F_1^* and F_2^* are distribution functions, H_0 can be reformulated as

$$H_0^* : F_0 = F_1^* + (1 - \omega^*)F_2^* \quad \text{properly} \qquad (5.4.4)$$

In other words, H_0^* is true if and only if H_0 is true.

Define

$$\alpha_{i0}^* = \int F_i^* \, dF_0 \ , \qquad i=1, 2 \qquad (5.4.5)$$

and

$$\alpha_{12}^* = \int F_1^* \, dF_2^* \ . \qquad (5.4.6)$$

From the assumption of continuity of F_1, F_2 and the definitions, F_1^* and F_2^* are continuous, hence we have

$$\alpha_{0i}^* = 1- \alpha_{i0}^* \ , \qquad \alpha_{21}^* = 1 - \alpha_{12}^*$$

Following the same argument as in section 3.1, under H_0^* we have a necessary condition

$$\alpha_{10}^* + \alpha_{02}^* - \alpha_{12}^* - \frac{1}{2} = 0 \ , \qquad (5.4.7)$$

By definition, $\alpha_{10}^* = b\alpha_{10} + (1- b)\alpha_{20}$, $\alpha_{20}^* = a\alpha_{10} + (1- a)\alpha_{20}$, and

$\alpha_{12}^{*} = \frac{1}{2}(1+ a -b) + (b- a)\alpha_{12}$, where α_{ij} is defined by (3.1.3), hence we may rewrite (5.4.7) as

$$0= \alpha_{10}^{*} + \alpha_{02}^{*} - \alpha_{12}^{*} - \frac{1}{2} = (b -a)(\alpha_{10} + \alpha_{02} - \alpha_{12} - \frac{1}{2}) \qquad (5.4.8)$$

Define

$$T_2^{*} = (b- a)T_2 = (b- a)(W_{10} + W_{02} - W_{12} - \frac{1}{2}),$$

$$T_3^{*} = (b- a)T_3 = (b- a)(W_{10} + R_{02} - R_{12} - \frac{1}{2}) , \qquad (5.4.9)$$

$$T_4^{*} = (b -a)T_4 = (b -a)(R_{10} + R_{02} - \alpha_{12} - \frac{1}{2}) ,$$

where W_{ij} is defined by (3.1.2) and R_{ij} by (3.3.1). The variances of T_2^{*}, T_3^{*}, T_4^{*} can be derived by using the variances of T_2, T_3, T_4 respectively. Similarly, consistent estimators of these variances can be derived from those of T_2, T_3 and T_4 by multiplying the factor $(b-a)^2$. Then large sample tests can formulated as follows:

(1) Problem C1

Reject H_0 if $|T_2^{*}|(\hat{\text{Var}}\ T_2^{*})^{-1/2} > z_{1- \alpha/2}$, (5.4.10)

(2) Problem C2

Reject H_0 if $|T_3^{*}|(\hat{\text{Var}}\ T_3^{*})^{-1/2} > z_{1-\alpha/2}$, (5.4.11)

(3) Problem C3

Reject H_0 if $|T_4^{*}|(\hat{\text{Var}}\ T_4^{*})^{-1/2} > z_{1-\alpha/2}$ (5.4.12)

where $z_{1- \alpha/2}$ is the $(1- \alpha/2)$x100 percentile of the standard normal distribution.

5.5 Problems D1, D2, D3

Suppose that each of F_0, F_1 is continuous and is either known, or there is a random sample from it, and assume that

$$F_0 = \omega F_1 + (1 - \omega)F_2 \text{ properly.} \qquad (5.5.1)$$

We wish to test the null hypothesis:

H_0: $F_2 = F$, a specified d. f.

Before discussing how to test H_0, let us first analyze the testing problem a little further. Given F_0 and F_1, does there exist a real number ω in [0,1] such that $F_0 - \omega F_1$ is a nondecreasing function? The answer to this question is contained in the following:

Proposition 5.1

Let F_0, F_1 be two cumulative distribution functions satisfying the condition that whenever $F_0(x') = F_0(x)$ for $x < x'$, it implies $F_1(x') = F_1(x)$. Define

$$G(x,\omega) = F_0(x) - \omega F_1(x) \tag{5.5.2}$$

then $G(x, \omega)$ is nondecreasing if and only if

$$\omega \leq \omega_0 \quad , \tag{5.5.3}$$

where $$\omega_0 = \inf\left\{\frac{F_0(x') - F_0(x)}{F_1(x') - F_1(x)} : x' > x, F_1(x') > F_1(x)\right\} \tag{5.5.4}$$

If furthermore, F_0, F_1 possess densities f_0, f_1 respectively, such that whenever $f_0(x) = 0$ it implies that $f_1(x) = 0$, then $G(x, \omega)$ is nondecreasing if and only if

$$\omega \leq \inf\left\{\frac{f_0(x)}{f_1(x)} : \quad f_1(x) > 0\right\} .$$

Remark: If F_0 is a proper mixture of F_1 and F_2, then the mixing proportion ω must satisfy

$$0 \leq \omega \leq \omega_0 .$$

On the other hand, by a similar argument, Proposition 5.1 can be applied to

$$H(x,\omega) = \omega F_0(x) - (1 - \omega) F_2(x)$$

and then ω must also satisfy

$$\omega \geq 1 - \omega_1 \quad ,$$

where

$$\omega_1 = \inf\left\{ \frac{F_0(x') - F_0(x)}{F_2(x') - F_2(x)} \; : \; F_2(x') > F_2(x),\ x' > x \right\} \tag{5.5.5}$$

In order that ω to be unique, we then must have

$$\omega_0 = 1 - \omega_1 \quad , \tag{5.5.7}$$

which can be used to derive test statistics for testing the hypothesis of mixture. (Not shown in this paper).

Proof:

Suppose that $G(x, \omega)$ is nondecreasing, then for $x < x'$ such that $F_1(x') > F_1(x)$, we have

$$\omega \leq \frac{F_0(x') - F_0(x)}{F_1(x') - F_1(x)} \quad .$$

It follows that, by taking infimum, $\omega \leq \omega_0$. Conversely if $\omega \leq \omega_0$, then for $x' > x$ and $F_1(x') > F_1(x)$, we have

$$\omega \leq \omega_0 \leq \frac{F_0(x') - F_0(x)}{F_1(x') - F_1(x)} \quad , \text{ or}$$

$$G(x', \omega) \geq G(x, \omega) \; .$$

For $x' > x$ and $F_1(x') = F_1(x)$, we have

$$G(X', \omega) - G(x,\omega) = F_0(x') - F_0(x) \geq 0 \quad .$$

If, furthermore, F_0, F_1 have densities f_0, f_1, respectively, then $G(x,\omega)$ is nondecreasing if and only if

$$f_0(x) - \omega f_1(x) \geq 0 \quad \text{for all } x \; .$$

The rest of the proposition follows. q. e. d.

From (5.5.1), F_2 can be expressed as

$$F_2 = (1-\omega)^{-1}(F_0 - \omega F_1) \qquad (5.5.8)$$

Conversely if there exists an ω in $(0,1)$ such that $F_0 - \omega F_1$ is nondecreasing, $(1-\omega)^{-1}(F_0 - \omega F_1)$ would be a distribution function and F_0 would be a mixture of F_1 and $(1-\omega)^{-1}(F_0 - \omega F_1)$. From Proposition 5.1 there might be infinitely many such , and the null hypothesis $H_0 : F_2 = F$ states that there is one ω which will make $(1-\omega)^{-1}(F_0 - \omega F_1)$ equal to F (a. s.)

Now under H_0, we have

$$F_0 = F_1 + (1-\omega)F \quad ,$$

and so we can use the statistics T_3 and T_4, discussed in sections 3.2, 3.3, to test H_0.

Another possible line to attack would be to use the following statistics as in Chapter IV:

$$\inf_{0\le\omega\le1} \sup_x | F_{0n_0}(x) - \omega F_{1n_1}(x) - (1-\omega)F(x) |, \qquad \text{(for problem D1)}$$

$$\inf_{0\le\omega\le1} \sup_x |F_{0n}(x) - \omega F_1(x) - (1-\omega)F(x)| \quad , \qquad \text{(for problem D2)}$$

$$\inf_{0\le\omega\le1} \sup_x |F_0(x) - \omega F_{1n}(x) - (1-\omega)F(x)| \quad , \qquad \text{(for problem D3)}$$

where $F_{in_i}(x)$ is the empirical distribution function.

5.6 Problem F1

In this section and the next two sections, we study problems of testing two mixtures simultaneously. Let $\{X_{0i}^{a}\}$ and $\{X_{0i}^{b}\}$ be

random samples from F_{a0} and F_{b0} with sample sizes n_{a0}, n_{b0} respectively. Assume that all the X's are mutually independent. Suppose that each of F_{a1}, F_{a2}, F_{b1}, F_{b2} is continuous and is either known, or there is a random sample from it. We wish to test the follwoing null hypothesis:

$$H_0: \quad F_{a0} = \omega_a F_{a1} + (1 - \omega_a)F_{a2} \qquad \text{properly and} \qquad (5.6.1)$$

$$F_{b0} = \omega_a F_{b1} + (1 - \omega_b)F_{b2} \qquad \text{properly.} \qquad (5.6.2)$$

For each single mixture we may use the test statistics T_2, T_3 or T_4 discussed in sections 3.1 - 3.3.

Since all the samples under consideration are assumed to be independent, we can combine the test statistics for each single mixture to form a test statistic for H_0 . In the following we use an example to illustrate how the statistic can be derived.

Suppose that F_{a1} is known and for each of F_{a2}, F_{b1}, F_{b2} there is a random sample $\{X_{2i}{}^a\}$, $\{X_{1i}{}^b\}$, $\{X_{2i}{}^b\}$ of size n_{a2}, n_{b1}, n_{b2} respectively. Also assume that all the samples are mutually independent. For testing (5.6.1) we use

$$T_3{}^a = R_{10}{}^a + W_{02}{}^a - R_{12}{}^a - \frac{1}{2} \qquad (5.6.3)$$

where $R_{ij}{}^a$ is similarly defined as in (3.2.1) and $W_{ij}{}^a$ as in (3.1.2) For testing (5.6.2) we use

$$T_2{}^b = W_{10}{}^b + W_{02}{}^b - W_{12}{}^b - \frac{1}{2} \qquad (5.6.4)$$

Var $T_3{}^a$, Var $T_2{}^b$ and their consistent estimators $\hat{\text{Var}}\ T_3{}^a$, $\hat{\text{Var}}\ T_2{}^b$ can be similarly derived as in sections 3.3, 3.2.

There are many possible methods of combining these two independent statistics, $T_3{}^a$, $T_2{}^b$, to test H_0. We will consider three of them:

(1) as all n's tend to ∞ ,

$$T_{10} = (T_3^a)^2(\hat{\text{Var}}\ T_3^a)^{-1} + (T_2^b)^2(\hat{\text{Var}}\ T_2^b)^{-1} \longrightarrow \chi_2^2 \qquad (5.6.5)$$

in distribution.

(2) Given observations $T_3^a = t_a$, $T_2^b = t_b$, define the following random variables:

$$-2 \log \Pr\{ T_3^a > t_a \} \quad ,$$

$$-2 \log \Pr\{ T_2^b > t_b \} \quad ,$$

then

$$T_{11} = -2 \log \Pr\{ T_3^a > t_a\} - 2\log\Pr\{ T_2^b > t_b\} \qquad (5.6.6)$$

will have approximately a chi-square distribution with four degree of freedom. (If T_3^a and T_2^b are continuous, this would be so exactly.)

(3) This will be discussed in section 5.7.

5.7 Problem F2

Let $\{X_{0i}^a\}$ and $\{X_{0i}^b\}$ be random samples form F_{a0} and F_{b0} with sizes n_{a0}, n_{b0} respectively. Assume that these two samples are independent. Suppose that each of F_{a1}, F_{a2}, F_{b2} is continuous and is either known or there is a random sample from it. We wish to test the following null hypothesis:

$$H_0: \quad F_{a0} = \omega_a F_{a1} + (1 - \omega_a)F_{a2} \qquad \text{properly} \qquad (5.7.1)$$

and

$$F_{b0} = \omega_b F_{a1} + (1 - \omega_b)F_{b2} \qquad \text{properly} \qquad (5.7.2)$$

The difference between (5.7.1) - (5.7.2) and (5.6.1)-(5.6.2) is that F_{a1} and F_{b1} are not necessarily identical in (5.6.1)-(5.6.2), but are identical in (5.7.1)-(5.7.2). If F_{a1} is known, the method discussed in section 5.7 can be applied here, but not if F_{a1} is only given by a

random sample. For in this latter case the test statistics for each single mixture are not necessarily (or usually) independent. Hence we have to investigate their covariance structure. On the other hand, the method given below can be applied to section 5.6 . (This is why it is included as method (3) at the end of that section.)

First suppose that there are random samples $\{X_{1i}^{a}\},\{ X_{2i}^{a}\} ,\{ X_{2i}^{b}\}$ of sizes n_{a1}, n_{a2}, n_{b2} from each of F_{a1}, F_{a2}, F_{b2} respectively. And assume that all samples are mutually independent. For testing (5.7.1) we use

$$T_2^{a} = W_{10}^{a} + W_{02}^{a} - W_{12}^{a} - \frac{1}{2} \tag{5.7.3}$$

where W_{ij}^{a} is similarly defined as in (3.1.2). For testing (5.7.2) we use

$$T_2^{b} = W_{10}^{b} + W_{02}^{b} - W_{12}^{b} - \frac{1}{2} \tag{5.7.4}$$

Then Var T_2^{a} and Var T_2^{b} can be derived by using (3.1.4), while

$$n_{a1}\ \mathrm{Cov}(T_2^{a}, T_2^{b}) = -(\alpha_{01}^{a} - \alpha_{21}^{a})(\alpha_{01}^{b} - \alpha_{21}^{b}) + \int [F_{a0} - F_{a2}][F_{b0} - F_{b2}]dF_{a1} \tag{5.7.5}$$

Define matrix

$$\Sigma_1 = \begin{pmatrix} \mathrm{Var}\ T_2^{a} & \mathrm{Cov}(T_2^{a}, T_2^{b}) \\ \mathrm{Cov}(T_2^{a}, T_2^{b}) & \mathrm{Var}\ T_2^{b} \end{pmatrix} \tag{5.7.6}$$

Let $N = n_{a0} + n_{a1} + n_{a2} + n_{b0} + n_{b2}$, and assume that as $N \longrightarrow \infty$, $n_{ij}/N \longrightarrow r_{ij}$ and that there exists ε such that $0 < \varepsilon \le r_{ij} \le 1-\varepsilon < 1$ for all n's. Then under H_0 , $N^{1/2}(T_2^{a}, T_2^{b})$ has an asymptotic bivariate normal distribution with mean vector (0, 0) and variance-covariance matrix $\lim_{N\to\infty} N\Sigma$.

Denote the consistent estimators of Var T_2^a, Var T_2^b, Cov(T_2^a, T_2^b), which can be derived as in section 3.1, by $\hat{\text{Var}}\ T_2^a$, $\hat{\text{Var}}\ T_2^b$, $\hat{\text{Cov}}(T_2^a, T_2^b)$ respectively, and introduce the matrix

$$S_1 = \begin{pmatrix} \hat{\text{Var}}\ T_2^a & \hat{\text{Cov}}(T_2^a, T_2^b) \\ \hat{\text{Cov}}(T_2^a, T_2^b) & \hat{\text{Var}}\ T_2^b \end{pmatrix} \tag{5.7.7}$$

Then by Theorems 2.1 and 2.2, the statistic

$$T_{12} = N(T_2^a, T_2^b)S_1^{-1}(T_2^a, T_2^b)' \tag{5.7.8}$$

has an asymptotic central chi-square distribution with 2 degrees of freedom under H_0. Generally it has a non-central chi-square distribution with 2 degrees of freedom, and noncentrality parameter

$$\lim_{N\to\infty} N(ET_2^a, ET_2^b)\ \Sigma_1^{-1}(ET_2^a, ET_2^b)' \quad , \tag{5.7.9}$$

where ET denotes the expectation of T.

Suppose that F_{a2} is known and there are (mutually independent) random samples from each of F's. Then, instead of T_2^a, we use

$$T_3^a = W_{10}^a + R_{02}^a - R_{12}^a - \frac{1}{2} \tag{5.7.10}$$

to test (5.7.1) and use

$$T_{13} = N(T_3^a, T_2^b)S_2^{-1}(T_3^a, T_2^b)' \tag{5.7.11}$$

to test H_0, where S_2 is similarly defined as S_1, i. e.

$$S_2 = \begin{pmatrix} \hat{\text{Var}}\ T_3^a & \hat{\text{Cov}}(T_3^a\ T_2^b) \\ \hat{\text{Cov}}(T_3^a, T_2^b) & \hat{\text{Var}}\ T_2^b \end{pmatrix} \tag{5.7.12}$$

5.8 Problem F3

Let $\{X_{0i}^{a}\}$ and $\{X_{0i}^{b}\}$ be random samples form F_{a0} and F_{b0} with sizes n_{a0}, n_{b0} respectively. Assume that these two samples are mutually independent. Suppose that each of F_{a1}, F_{a2} is continuous and is either known or there is a random sample from it. We wish to test the following null hypothesis:

$$H_0: \quad F_{a0} = \omega_a F_{a1} + (1 - \omega_a)F_{a2} \quad \text{properly} \tag{5.8.1}$$

and

$$F_{b0} = \omega_b F_{a1} + (1-\omega_b)F_{a2} \quad \text{properly} . \tag{5.8.2}$$

Note that the difference between (5.6.1), (5.6.2), (5.7.1), (5.7.2) and (5.8.1), (5.8.2) is that $F_{a1} = F_{b1}$ and $F_{a2} = F_{b2}$ in (5.8.1), (5.8.2). If both F_{a1} and F_{a2} are known, we may apply methods (1) and (2) discussed in section 5.6. If only one of F_{a1} and F_{a2} is known while there is a random sample for the other, we may apply the method discussed in section 5.7. Now suppose that for each of F_{a1}, F_{a2}, there is a random sample $\{X_{1i}^{a}\}$, $\{X_{2i}^{a}\}$ with size n_{a1}, n_{a2} respectively. Also assume that all samples are mutually independent. For testing (5.8.1) we use

$$T_2^{a} = W_{10}^{a} + W_{02}^{a} - W_{12} - \frac{1}{2} , \tag{5.8.3}$$

where W_{ij}^{a} is similarly defined as in (3.1.2). For testing (5.8.2), we use

$$T_2^{b} = W_{10}^{b} + W_{02}^{b} - W_{12} - \frac{1}{2} \tag{5.8.4}$$

Then Var T_2^{a}, Var T_2^{b} can be derived by using (3.1.4), while

$$\text{Cov}\ (T_2^{a},\ T_2^{b})$$

$$= \frac{1}{n_{a1}}\left\{ -(\alpha_{01}^{a} - \alpha_{21})(\alpha_{01}^{b} - \alpha_{21}) + \int (F_{a0} - F_{a2})(F_{b0} - F_{a2})dF_{a1}\right\}$$

$$+ \frac{1}{n_{a2}}\left\{-(\alpha_{02}^{a} - \alpha_{12})(\alpha_{02}^{b} - \alpha_{12}) + \int (F_{a0} - F_{a1})(F_{b0} - F_{b1})dF_{a2}\right\}$$

$$- \frac{1}{n_{a1}n_{a2}}\left\{ -\alpha_{12}^{2} - \alpha_{21}^{2} - \alpha_{12}\alpha_{21} + \int F_{a1}^{2}dF_{a2} + \int F_{a2}^{2}dF_{a1}\right\} \tag{5.8.5}$$

Define the matrix

$$\Sigma_3 = \begin{pmatrix} \text{Var}\ T_2^{a} & \text{Cov}(T_2^{a},\ T_2^{b}) \\ \text{Cov}(T_2^{a},\ T_2^{b}) & \text{Var}\ T_2^{b} \end{pmatrix} \tag{5.8.6}$$

and denote the corresponding consistent estimator of Σ_3 by S_3.

Let $N = n_{a0} + n_{b0} + n_{a1} + n_{a2}$, and assume that as $N \longrightarrow \infty$, $n_{ij}/N \longrightarrow r_{ij}$ and there exists ε such that $0 < \varepsilon \le r_{ij} \le 1-\varepsilon < 1$, for all n's. Then under H_0, by Theorems 3.1 and 3.2, the statistic

$$T_{14} = N(T_2^{a},\ T_2^{b})S_3^{-1}(T_2^{a},\ T_2^{b})' \tag{5.8.7}$$

has an asymptotic central chi-square distribution with two degrees of freedom. In general T_{14} has an asymptotic non-central chi-square distribution with two degrees of freedom and noncentrality parameter

$$\lim_{N\to\infty} N(ET_2^{a},\ ET_2^{b})\Sigma_3^{-1}(ET_2^{a},\ ET_2^{b})' \ . \tag{5.8.8}$$

BIBLIOGRAPHY

1. Behboodian, F. (1972), "On the distribution of a symmetric statistic from a mixed population," Technometrics, 14, 919-923.

2. Behboodian, J. (1976), "Estimation of the parameters of finite location and scale mixtures," Technical Report No. 28, Dept. Statistics, Stanford University.

3. Birnbaum, Z. W. and Klose, O. M. (1957), "Bounds for the variance of the Mann-Whiteney statistic," Ann. Math. Statist., 28, 933-945.

4. Birnbaum, Z. W. and McCarthy, R. C. (1958), "A distribution -free upper confidence bound for $P\{Y < X\}$, based on independent samples of X and Y," Ann. Math. Statist., 29, 558-562.

5. Breiman, L. (1968), Probability. Addison-Wesley Publishing Co., Reading, Mass..

6. Bryant, P. (1973), "The sample covariance matrix in mixture problems: Part I: Scatterplots and estimators, Part II: Tests of hypotheses," Technical Report No. 320-2084, IBM Data Processing Division, Cambridge Scientific Center, Cambridge, MA 02139.

7. Cramer, H. (1946), Mathematical Methods of Statistics, Princeton University Press, Princeton, N. J..

8. David, F. N., Kendall, M. G. and Barton, D. E. (1966), Symmetric Function and Allied Tables, University Press, Cambridge, England.

9. Goodwin, E. T. (1949), "The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(x)e^{-x^2}dx$," Proc. Cambridge Phil. Soc., 45, 241-245.

10. Hariton, G. L. (1972), Multivariate Mixed Models, Ph. D. dissertation, Dept. Math., University of Toronto.

11. Johnson, N. L. (1973), "Some simple tests of mixtures with symmetric components," Comm. Statist., 1, 17-25.

12. Lehmann, E. L. (1959), Testing Statistical Hypotheses, John Wiley & Sons, Inc., New York.

13. Puri, M. L. and Sen, P. K. (1971), Nonparametric Methods in Multivariate Analysis, John Wiley & Sons, Inc., New York.

14. Thomas, E. A. C. (1969), "Distribution-free tests for mixed probability distributions," Biometrika, 56, 475-484.

15. Woinsky, M. N. and Kurz, L. (1969), "Sequential non-parametric two-way classification with prescribed maximum asymptotic error probability," Ann. Math. Statist., 40, 445-455.